Evolutionary dynamics of intergenic sites, pan-genomes, and introgression in bacteria

Thorpe, Harry

Award date:
2018

Awarding institution:
University of Bath

Evolutionary dynamics of intergenic sites, pan-genomes, and introgression in bacteria

Harry Arthur Frank Wright Thorpe

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

September 2017

Table of contents

Acknowledgements **2**

List of publications **3**

Declarations **4**

Abstract **5**

Chapter 1: Introduction **6**

Chapter 2: Comparative analyses of selection operating on non-translated intergenic regions of diverse bacterial species **37**

Chapter 3: Piggy: A Rapid, Large-Scale Pan-Genome Analysis Tool for Intergenic Regions in Bacteria **62**

Chapter 4: Variation in deleterious mutation load in *H. pylori* populations, and effect of selection on introgressed DNA in hpEurope **80**

Chapter 5: Compensatory evolution is widespread in Rho-independent terminators in bacteria **95**

Chapter 6: Discussion **108**

Appendix **116**

References **139**

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Ed Feil. He has given me guidance, the opportunity to travel and collaborate with others, and the freedom to pursue my own ideas in the wonderful world of bacterial evolution. Ed has always been extremely generous with his time and knowledge, and I have learnt more during these four years than I could have possibly imagined when I started.

I would also like to thank Laurence Hurst for valuable input on the work in chapter 2. I would like to thank Daniel Falush for introducing me to the world of *Helicobacter pylori*, and to Kaisa Thorell and Koji Yahara, with whom I collaborated for the work in chapter 3. I would like to thank Alan McNally for encouragement and feedback on early versions of Piggy. I would like to thank Nicola Coyle for providing the *V. anguillarum* and *V. parahaemolyticus* data used in chapter 5. I would like to thank Sion Bayliss for the many fruitful discussions, collaborations, and ongoing work which did not make it into this thesis.

I would like to thank my parents for all the opportunities they have given me, and for encouraging me to follow my interests. Finally, I would like to thank Anushka for her love and support, and for making my world a brighter place.

List of publications

The following peer-reviewed publication and preprint correspond to work which directly contributed to this thesis:

Thorpe, H.A., Bayliss, S.C., Hurst, L.D., Feil, E.J., 2017. Comparative Analyses of Selection Operating on Non-translated Intergenic Regions of Diverse Bacterial Species. Genetics 206, 363–376.

Thorpe, H.A., Bayliss, S.C., Sheppard, S.K., Feil, E.J., 2017. Piggy: A Rapid, Large-Scale Pan-Genome Analysis Tool for Intergenic Regions in Bacteria. bioRxiv.

The following peer-reviewed publications correspond to collaborations on work which did not directly contribute to this thesis:

Bayliss, S.C., Hunt, V.L., Yokoyama, M., **Thorpe, H.A.**, Feil, E.J., 2017. The use of Oxford Nanopore native barcoding for complete genome assembly. Gigascience.

Senghore, M., Bayliss, S.C., Kwambana-Adams, B.A., Foster-Nyarko, E., Manneh, J., Dione, M., Badji, H., Ebruke, C., Doughty, E.L., **Thorpe, H.A.**, Jasinska, A.J., Schmitt, C.A., Cramer, J.D., Turner, T.R., Weinstock, G., Freimer, N.B., Pallen, M.J., Feil, E.J., Antonio, M., 2016. Transmission of Staphylococcus aureus from Humans to Green Monkeys in The Gambia as Revealed by Whole-Genome Sequencing. Appl. Environ. Microbiol. 82, 5910–5917.

Senn, L., Clerc, O., Zanetti, G., Basset, P., Prod'hom, G., Gordon, N.C., Sheppard, A.E., Crook, D.W., James, R., **Thorpe, H.A.**, Feil, E.J., Blanc, D.S., 2016. The Stealthy Superbug: the Role of Asymptomatic Enteric Carriage in Maintaining a Long-Term Hospital Outbreak of ST228 Methicillin-Resistant Staphylococcus aureus. MBio 7, e02039–15.

Laabei, M., Uhlemann, A.-C., Lowy, F.D., Austin, E.D., Yokoyama, M., Ouadi, K., Feil, E., **Thorpe, H.A.**, Williams, B., Perkins, M., Peacock, S.J., Clarke, S.R., Dordel, J., Holden, M., Votintseva, A.A., Bowden, R., Crook, D.W., Young, B.C., Wilson, D.J., Recker, M., Massey, R.C., 2015. Evolutionary Trade-Offs Underlie the Multi-faceted Virulence of Staphylococcus aureus. PLoS Biol. 13, e1002229.

Declaration

I confirm that the work presented in this thesis is the work of the author. Due to the highly collaborative nature of science, some small data preparation steps and analyses were performed by collaborators. Where this is the case, it is clearly denoted in square brackets within the text, for example: [N.B. This analysis was performed by XXXX, XXXX university.]. Otherwise the work is the sole work of the author.

Chapters 2 and 3 are written in the style of peer-reviewed publications (chapter 2 is published, and chapter 3 is under review). In these chapters ‘we’ is used instead of ‘I’, and a statement of authorship is provided for these chapters in the Appendix.

Abstract

Bacterial genomes evolve under strong selective constraints. They are compact, gene dense, and contain little extraneous DNA. They are also diverse; individuals from the same species frequently differ substantially in gene content from each other. The work presented in this thesis has investigated the diversity of bacterial species and the selective forces which shape their genomes. A focus has been given to regions of the genome which are poorly understood, particularly intergenic sites. Intergenic sites are shown to be under widespread purifying selection which varies according to divergence time, the class of regulatory element, and distance from gene borders. This is complemented by work on Rho-independent terminator sequences which shows that compensatory evolution is widespread in many species. A detailed analysis of *H. pylori* introgression shows that selection acts to moderate the uptake of DNA from different sources. Additionally, analyses of pan-genomes incorporating intergenic regions were performed, and a new tool, Piggy, was introduced to facilitate these analyses. This enabled the interaction between genes and their cognate intergenic regions to be analysed, and genes with divergent upstream intergenic regions were shown to be more differentially expressed than those without in *S. aureus*. This work has provided new insights into the evolutionary dynamics of these poorly understood but vital components of bacterial genomes.

Chapter 1

Introduction

Introduction

Bacteria are ubiquitous, ecologically important organisms which affect almost every conceivable part of the biosphere. Recent advances in whole-genome sequencing technologies have provided great insight into their genomic composition and diversity. The work in this thesis uses large-scale sequencing data to investigate the evolutionary dynamics of parts of bacterial genomes which are poorly understood, primarily intergenic regions. As each chapter has a focused introduction, in this overall introduction I provide a general introduction to bacterial evolutionary genomics, from the basic composition of the genome to current understanding of the forces which shape it.

The structure and organisation of bacterial genomes

Bacterial genomes vary in size by more than two orders of magnitude, from the tiny *Nasuia*-ALF genome (112 Kb) encoding only 137 genes (Bennett and Moran 2013), to the huge *Sorangium cellulosum* genome (14.8 Mb) encoding 11,599 genes (Han et al. 2013). These are extreme examples, and bacterial genomes typically range from 1.5-8 Mb in length. Based on an analysis of 659 sequenced bacterial genomes, Koonin and Wolf suggested that genome size in bacteria is bimodal, with one peak at 5 Mb and another at 2 Mb (Koonin and Wolf 2008). A recent reanalysis of this distribution with 3923 bacterial genomes also found a bimodal distribution, however the distribution became unimodal as redundant species were removed (Gweon, Bailey, and Read 2017) (Figure 1.1e). Thus, the bimodal distribution is likely an artefact of sequencing bias towards species of interest; for example *Escherichia coli* and *Salmonella enterica* in the 5 Mb peak and *Helicobacter pylori* and *Staphylococcus aureus* in the 2 Mb peak. After this bias is accounted for, a unimodal distribution with a peak of approximately 3 Mb is found (Gweon, Bailey, and Read 2017).

The majority of bacteria have a single circular chromosome, which is replicated bidirectionally from the origin of replication to the terminus (Figure 1.1a). However, bacteria can vary in genome organisation with some species having multiple circular and/or linear chromosomes (Egan, Fogel, and Waldor 2005) (Figure 1.1b). This was first shown in *Rhodobacter sphaeroide*, where Pulsed-Field Gel Electrophoresis (PFGE) indicated the presence of two circular chromosomes (Suwanto and Kaplan 1989). Other species with multiple circular chromosomes include members of the *Vibrio* genus (two chromosomes), and *Burkholderia cenocepacia* (three chromosomes) (Cooper et al. 2010). *Borrelia burgdorferi* has a linear chromosome and multiple

linear and circular plasmids (Fraser et al. 1997), and *Agrobacterium tumefaciens* has both circular and linear chromosomes (Allardet-Servent et al. 1993). Bacteria also carry plasmids, which are semi-autonomous replicating units of DNA (Figure 1.1c). Plasmids are frequently transferred between species through conjugation, and often carry genes encoding selectively important traits (such as antibiotic resistance genes) (Harrison and Brockhurst 2012). Secondary chromosomes can be small and plasmids can be large, so differentiating them is not always clear, but typically chromosomes are essential and plasmids are dispensable (Egan, Fogel, and Waldor 2005). However, essentiality is not easy to measure, as it is context dependent. A secondary chromosome may be dispensable in rich laboratory media but not in the wild, and a plasmid may be essential in the presence of an antibiotic. Replication of chromosomes is also linked to the cell cycle, whereas replication of plasmids is not (Egan, Fogel, and Waldor 2005). To summarise, bacteria carry their genome in one circular chromosome (usually) or a small number of circular and linear chromosomes (less commonly), and a number of small plasmids (Figure 1.1a-c).

Bacterial genomes are compact, densely packed with genes, and highly organised. They have few repetitive regions and little extraneous DNA. This genome structure contrasts sharply with eukaryotic genomes, which are large, sparsely populated with genes, and repeat dense. To put this into context, on average 87% of bacterial genomes is protein coding (McCutcheon and Moran 2011), compared to 1.1% in humans (Venter et al. 2001). Genes in bacteria are commonly grouped together in operons, which are groups of co-transcribed genes (Jacob and Monod 1961) (Figure 1.1d). Organising genes as operons provides an elegant way to co-regulate genes. The length distribution of bacterial genes is relatively narrow, with most genes being approximately 1000 bp in length, a figure which is consistent across a broad range of taxa (Koonin and Wolf 2008) (Figure 1.1f). They are typically present as uninterrupted open reading frames (ORFs). The length distribution of intergenic regions is bimodal, with a peak at ~0 bp and another peak at ~150 bp (Figure 1.1f). The peak at 0 likely corresponds to small within-operon intergenic regions, and the peak at 150 bp likely corresponds to the intergenic regions located between operons (Koonin and Wolf 2008). There is a fat tail of longer intergenic regions; these likely correspond to regulatory elements, non-coding RNAs, and possibly unannotated genes. The between-operon intergenic regions harbour regulatory elements which are used to control gene expression. These include, but are not limited to promoters,

rho-dependent and independent transcriptional terminators, regulator binding sites, non-coding RNAs, and ribosome binding sites.

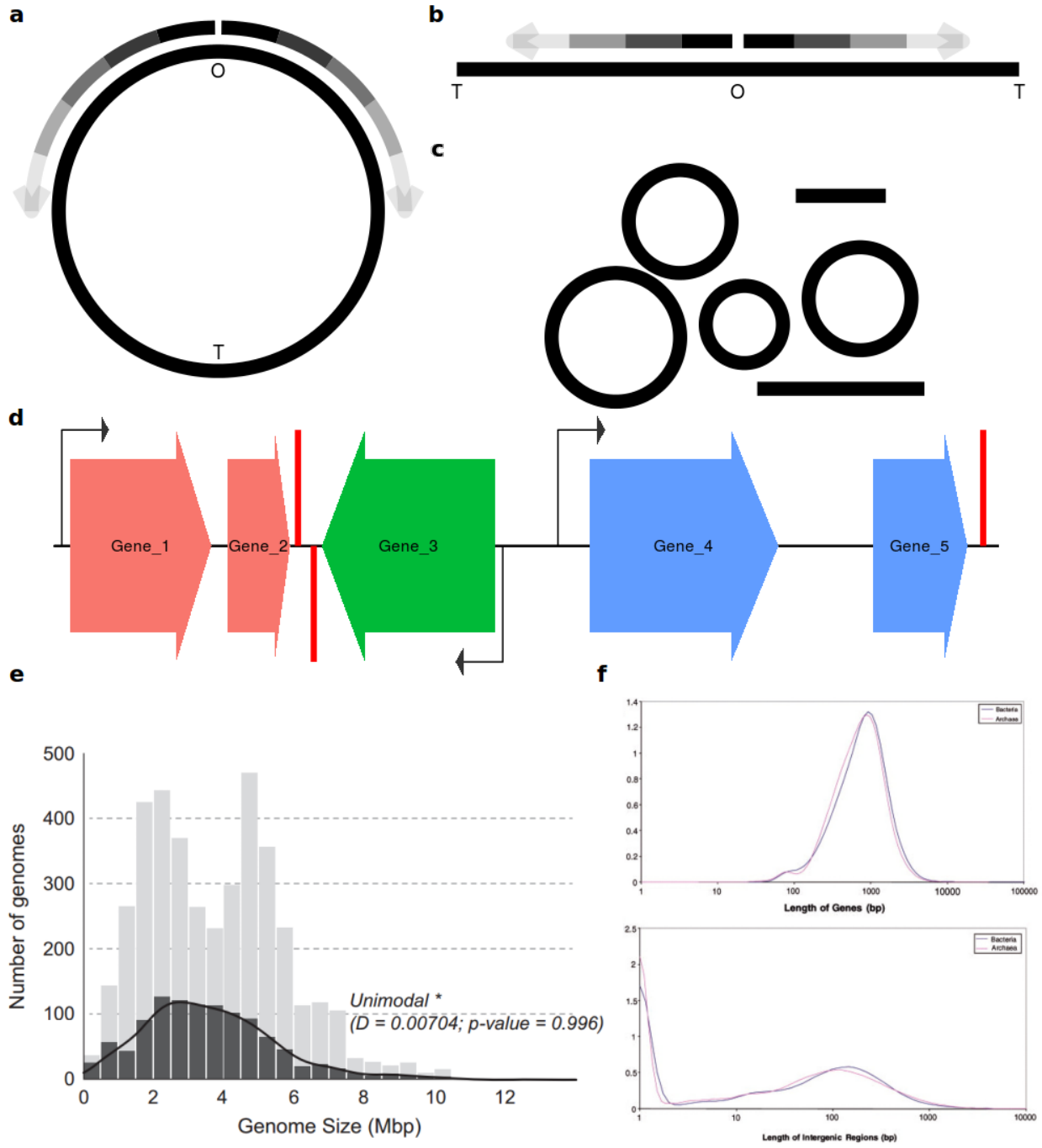

Figure 1.1: The composition of bacterial genomes. a. Circular and **b.** linear bacterial chromosomes, with the origin and terminus of replication marked. Replication proceeds bidirectionally from the origin to the terminus, as shown by the arrows. **c.** Circular and linear plasmids. **d.** A schematic of part of a bacterial genome, showing genes (coloured block arrows), promoters (angled black arrows), and terminators (red bars). Genes are coloured according to their operon. **e.** Size distribution of bacterial genomes, with all genomes (light grey), and redundant genomes removed (dark grey). **f.** Size distributions of genes (top), and intergenic regions (bottom). Figure adapted from (Gweon, Bailey, and Read 2017; Koonin and Wolf 2008).

Bacterial gene expression and regulation

The paradigm of bacterial gene regulation is the operon (Koonin and Wolf 2008; Jacob and Monod 1961). Operons typically consist of 2 to 4 genes which are co-transcribed and translated, enabling efficient expression of genes which are required to be expressed together (Figure 1.1d, Figure 1.2a). An operon will typically have a single promoter upstream of the first gene, and a terminator downstream of the last gene, and may have binding sites for regulators. The *lac* operon enables *E. coli* to metabolise lactose, and is a classic example of a typical bacterial operon (Jacob and Monod 1961; Lewis et al. 1996) (Figure 1.2a). The *lac* operon consists of three operational genes, *lacZ, lacY,* and *lacA*, which are sufficient for breaking down lactose. Upstream of *lacZ* is a promoter, and downstream of *lacA* is a terminator. Additionally, there is an operator located between the promoter and *lacZ*; this is a binding site for the *lac* repressor *lacI*, which is encoded on a separate operon (Jacob and Monod 1961; Lewis et al. 1996). These components enable the *lac* operon to be regulated according to environmental conditions. In the absence of lactose, the *lac* repressor binds the operator, preventing transcription of the *lac* operon. In the presence of lactose, lactose allosterically binds the *lac* repressor, preventing it from binding the operator. With the operator unbound, the promoter is free and transcription initiation can begin.

Bacterial transcription initiation is achieved through a coordinated effort involving RNA polymerase, sigma factors, promoters, and sometimes other regulatory proteins. RNA polymerase is a protein consisting of four subunits, β and β′ which form the active site, α2 which binds certain promoters, and ω which assists the folding of the β′ subunit (Figure 1.2c). RNA polymerase alone can perform transcription, but cannot bind the promoter, and so cannot initiate transcription. Transcription initiation is performed with the assistance of sigma factors (Browning and Busby 2004).

Sigma factors are proteins which bind RNA polymerase and enable it to bind the promoter. They consist of up to four domains, and these domains bind to specific promoter sequences (Browning and Busby 2016). There are a variety of sigma factors which bind to specific subsets of promoters, and these are used to globally regulate transcription. All bacteria have a dominant housekeeping sigma factor (such as σ70 in *E. coli*), which is involved in the expression of the majority of genes. These sigma factors have four domains, all of which are responsible for binding a specific part of the promoter (Browning and Busby 2016) (Figure 1.2b). Bacteria also

have alternative sigma factors, which are often evolutionarily related to the housekeeping sigma factors, but do not necessarily share the same four domain structure. Additionally, there are a class of alternative sigma factors (such as *E. coli* σ54) which are evolutionarily unrelated to other sigma factors, and bind different promoter motifs (Browning and Busby 2016).

Bacterial promoters consist of short sequence motifs which are recognised by sigma factors (Figure 1.2b). Promoters which are recognised by the σ70 family of sigma factors have -10 and -35 elements, which are 6 bp motifs which bind to subunits 2 and 4 of the sigma factor, respectively. The consensus sequences of the motifs are TATAAT for the -10 element and TTGACA for the -35 element (Browning and Busby 2016). Of these two motifs, the -10 element is more important for transcription initiation than the -35 element, although both are usually present. More recent work has shown the existence of other promoter elements, the extended -10 element (consensus sequence TGTG), and the discriminator element (consensus sequence GGG) (Hook-Barnard and Hinton 2007; Browning and Busby 2016). Some promoters also have an UP element (consensus sequence AAAWWTWTTTTNNNAAANN), which is bound by the C-terminus of the α subunit of RNA polymerase. Promoters bound by the σ54 family of sigma factors have a different structure consisting of -12 and -24 elements instead of -10 and -35 elements (Wigneshweraraj et al. 2008).

The modular nature of bacterial promoters enables their strength to be tailored; this provides a mechanism for controlling gene expression. It has been noted that no naturally occurring promoters show perfect matches to the consensus sequence for each element, and the different elements often vary in how much they deviate from the consensus (Browning and Busby 2016; Hook-Barnard and Hinton 2007). The degree of deviation from the consensus sequence influences the binding strength of the promoter to the sigma factor (a perfect consensus element binds more strongly than one which deviates from the consensus). This means that genes controlled by promoters with near-consensus sequences will be more highly expressed than those controlled by promoters which deviate further from the consensus. This variability in promoter strength provides a ‘baseline’ level of gene expression, where some genes can be constitutively more highly expressed than others (Browning and Busby 2016; Hook-Barnard and Hinton 2007; Hawley and McClure 1983).

Bacteria use two methods to terminate transcription: Rho-independent (or 'intrinsic') termination, and Rho-dependent termination (Peters, Vangeloff, and Landick 2011; Santangelo and Artsimovitch 2011). Rho-independent termination relies on intrinsic properties of the mRNA transcript in order to halt transcription. These properties consist of a GC rich stem-loop (hairpin) structure followed by a U-rich tract of the mRNA (Figure 1.2d). The stem has a mean length of 8 bp, compared to 4 bp for the loop, and 7-8 bp for the U-tract. Rho-independent termination occurs in four stages (of which the first three are shown in Figure 1.2e). First, the U-tract induces a pause in transcription. Second, the hairpin nucleates, meaning that the complementary bases bind each other to begin to form the stem-loop structure. Third, the hairpin extends; this melts ~3 bp of the RNA-DNA hybrid by pulling the RNA strand away from the hybrid. Fourth, RNA polymerase dissociates from the DNA, terminating transcription (Peters, Vangeloff, and Landick 2011). The strong sequence properties of Rho-independent terminators means they can be accurately predicted from the genome sequence. De Hoors *et al.* developed a decision rule to predict Rho-independent terminators with 94% specificity across 57 Firmicute species (De Hoors et al. 2005). Subsequently, Kingsford *et al.* developed a program, TransTermHP, which is several orders of magnitude faster than the decision rule and almost as accurate (Kingsford, Ayanbule, and Salzberg 2007).

In contrast, Rho-dependent termination does not have a simple set of intrinsic motifs, and instead relies on the presence of auxiliary proteins to complete termination. Indeed, the only common feature of Rho-dependent terminators is a richness of C residues in the mRNA transcript (Ciampi 2006). This means they cannot be predicted computationally from the genome sequence alone. There are a number of trans-acting factors which are required for transcription termination. Rho is a hexameric protein which is sufficient to terminate transcription at most Rho-dependent terminators in-vitro. It is essential in *E. coli*, and Rho homologs are ubiquitous across bacteria (Ciampi 2006). Rho acts as an ATP dependent RNA-DNA helicase, and moves along the mRNA until it encounters the transcription elongation complex stalled at a pause site. Transcription is then terminated by disruption of the RNA-DNA hybrid. In vivo, another protein, NusG is required in addition to Rho at certain terminators (Ciampi 2006).

Bacteria vary in how much they rely on the two forms of termination. In the Firmicutes, Rho-independent termination is the prefered method (Kingsford, Ayanbule, and Salzberg 2007; Ciampi 2006). Consistent with this, in *B. subtilis* Rho is not essential. In contrast, *E. coli* uses

Rho-independent and dependent termination approximately equally, consistent with Rho being essential in this species (Ciampi 2006).

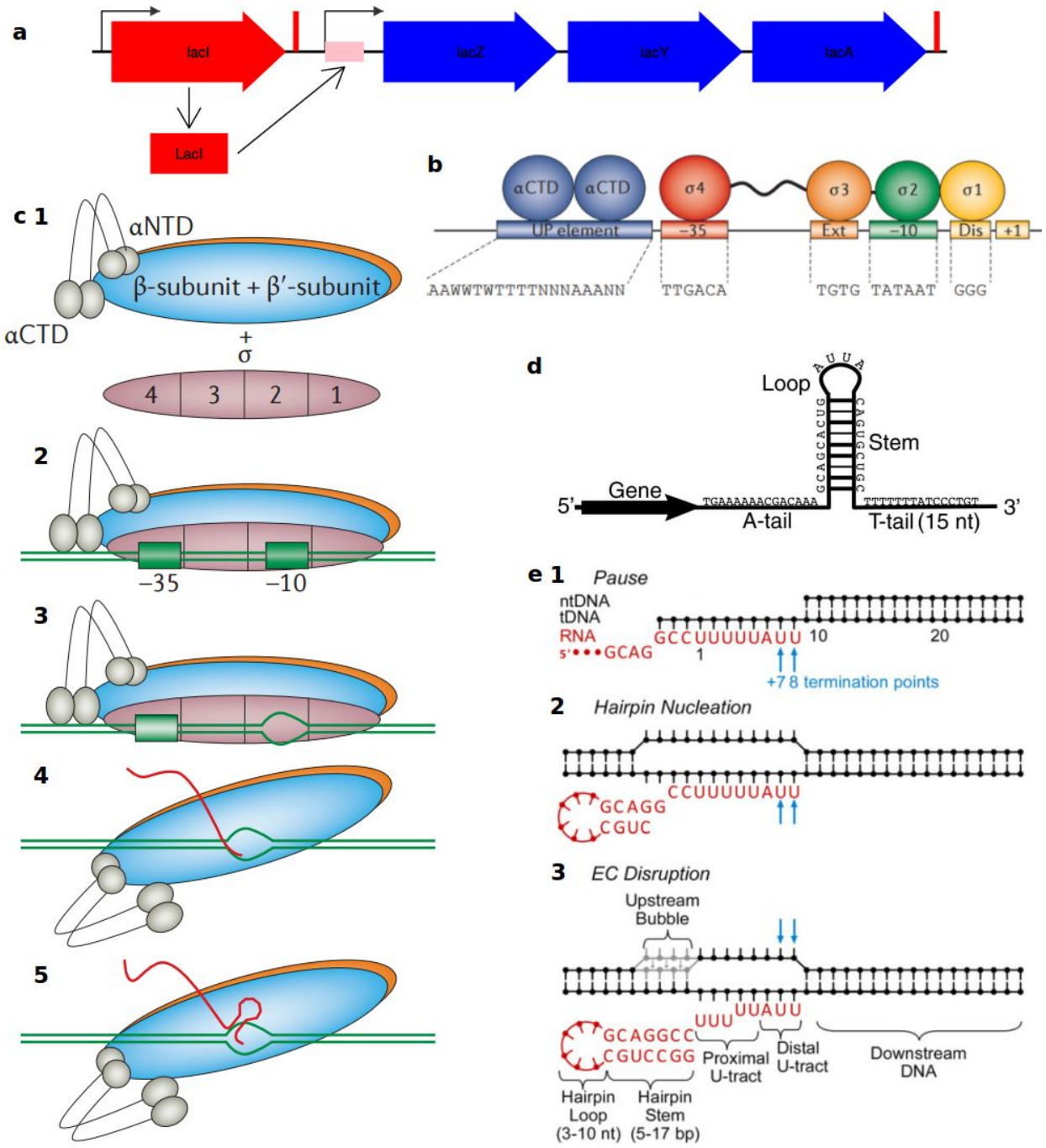

Figure 1.2: Bacterial operons, transcription initiation, and termination. a. Schematic of the *lac* operon. **b.** A σ-like sigma factor bound to a promoter. The αCTD domains of RNA polymerase are also shown. **c.** The components required for and mechanism of transcription initiation in bacteria. **d.** The structure of Rho-independent terminators. **e.** The mechanism of Rho-independent transcriptional termination. Figure adapted from (Browning and Busby 2016; Kingsford, Ayanbule, and Salzberg 2007; Peters, Vangeloff, and Landick 2011).

Bacterial typing methods

Within a bacterial species, individual strains vary considerably in phenotypes such as virulence and antibiotic resistance. In order to understand this variation, and to provide neutral markers for molecular epidemiology, accurate typing methods are needed. Early typing methods such as Pulsed-Field Gel Electrophoresis (PFGE) and Arbitrarily Primed Polymerase Chain Reaction (AP-PCR) both relied on observing banding patterns of DNA fragments on agarose gels (Macfarlane et al. 1999; Roberts et al. 1998). In PFGE, the genomic DNA is fragmented with rare cutting restriction enzymes, whereas in AP-PCR, a single random primer is used to amplify genomic DNA. Both methods detect variation which evolves quickly and therefore have good discriminatory power. This has enabled outbreak cases to be discriminated from the background population (Macfarlane et al. 1999; Roberts et al. 1998). In contrast Multi Locus Enzyme Electrophoresis (MLEE) detects more slowly evolving variation, and so is more useful for studying bacterial population structure over longer time scales (Selander et al. 1986; Milkman 1973). MLEE works by analysing the electrophoretic mobilities of housekeeping enzymes on starch gels, where molecular variation in protein sequences can be assayed and combined in order to produce a relatedness matrix of the bacterial strains. The crucial difference between [PFGE, AP-PCR] and MLEE is that the former is an unbiased assessment of genetic variation *en masse*, while in the latter only a small subset of enzymes common to the strains are analysed. This feature of MLEE meant that it could be used for population genetic analyses, and for the first time this enabled a view into the population structure of several bacterial species (Enright and Spratt 1999; J. M. Smith et al. 1993; Milkman 1973).

However, the methods discussed above suffered from several problems. They were slow and laborious, and samples could not easily be compared between laboratories. In PFGE and AP-PCR, the loci responsible for the variation were unknown, and in all three methods between-species comparisons were not possible (Enright and Spratt 1999). To address these issues MultiLocus Sequence Typing (MLST) was developed, based on the principles of MLEE (M. C. Maiden et al. 1998). In MLST, a set of housekeeping genes (usually 7) are chosen, and from these primers are designed to amplify a portion (~450 bp) of each gene. These PCR fragments are then sequenced, and the sequence at each gene is converted into an allele number. If the allele has been sequenced previously, it is assigned the existing allele number, and if it is novel then it is assigned a new number. The allele numbers at each gene can then be used to give an allele profile for the strain (Enright and Spratt 1999; M. C. Maiden et al. 1998).

Because sequence data are truly portable between laboratories, datasets could be combined and analysed together for the first time. This was made possible by the development of a website, pubmlst (https://pubmlst.org/), which stored the details of the primer sets for each organism. Further, pubmlst housed databases of the allele profiles of strains provided by different laboratories. At the time of writing, pubmlst has >200,000 isolates from 95 bacterial species. In order to understand and visualise the relationships between different allele profiles, an algorithm called eBURST was developed. eBURST aimed to classify allele profiles into related 'clonal complexes', each centred on the likely founding genotype (Feil et al. 2004). An example of an eBURST diagram is shown in Figure 1.3a.

MLST has provided great insight into bacterial population structure and evolutionary processes, but it has limited resolution as only 7 genes are analysed (<1% of the genome) (M. C. J. Maiden et al. 2013; Feil 2015). Since the genome of *Haemophilus influenzae* was fully sequenced in 1995 (Fleischmann et al. 1995), improvements in sequencing technologies and falling prices mean that it is now routine to sequence a bacterial genome for as little as £50. The research community has quickly taken advantage of these developments, and there are now >400,000 sequenced bacterial genomes in the sequence read archive (SRA), with the most important pathogens being represented by 1000s of isolates. This has provided unprecedented power to study evolutionary processes, as in principle all genetic variation is covered by whole-genome sequencing. However, the increased resolution offered by whole-genome sequencing has posed new challenges for data analysis. One solution which is in keeping with traditional MLST is to extend the scheme to include thousands of core alleles; this is known as cgMLST (M. C. J. Maiden et al. 2013). Alternatively the need to group isolates into 'types' can be removed, and the relationships between isolates can be interpreted by more continuous measures, such as distances on a phylogenetic tree or their shared accessory gene content (Feil 2015).

Bacterial population structure

Bacteria are asexual and reproduce by binary fission. Thus, in the absence of recombination to shuffle variation within the population, they evolve clonally and there is strong linkage disequilibrium between loci. If the rate of recombination is sufficiently high, then variation is shuffled and linkage disequilibrium is weak. In a landmark paper, Maynard Smith *et al.* analysed MLEE data from a number of bacterial species (J. M. Smith et al. 1993). It was found that

bacteria vary greatly in their population structure, from the highly clonal *Salmonella enterica* to the effectively panmictic (random associations between loci) *Neisseria gonorrhoeae*. However, two intermediate population structures were also found. *Neisseria meningitidis* displayed an 'epidemic' population structure, where associations between loci were found in a small number of common electrophoretic types. When these were removed, the associations between loci were no longer present. This population structure results from a rapid expansion of a small number of electrophoretic types. In contrast, *Rhizobium meliloti* exhibits associations between loci, but these are a product of the population being split into two groups, where there are random associations between loci within each group but not between the groups (J. M. Smith et al. 1993). This study revealed that the nature of bacterial population structure is highly variable between species.

Although the study described above provided great insight into the features of bacterial evolution, the nature of MLEE data prevents the electrophoretic types from being robustly classified, and comparisons between species cannot be made quantitatively. The eBURST algorithm developed by Feil *et al.* aimed to robustly classify groups of closely related strains, and to understand how they are evolutionarily related to each other (Feil et al. 2004). According to eBURST, strains which share the same allele profile are designated the same sequence type (ST). The STs are then grouped by thresholds, for example all STs in a group must share a minimum number of alleles with at least one other member of the group (Feil et al. 2004). If this group of STs is sufficiently stringent (only 1/7 alleles can vary between members), then it is known as a clonal complex (CC), and strains which vary at a single locus are known as single locus variants (SLVs). The founding member of the CC is then predicted by parsimony as the strain with the most SLVs (Feil et al. 2004). This was the first species-independent, sequence-based, and quantitative method to classify bacteria, and it enabled powerful snapshots of the population structure of different bacterial species to be compared with each other.

eBURST revealed that, in general, bacteria display a population structure consisting of multiple distantly related clones, each with recent clonally related descendents. However, the nature and robustness of the identified clonal complexes varies by species. In *S. aureus*, which diversifies primarily through mutation (rather than recombination) (Feil et al. 2003), the clonal complexes are robust (Figure 1.3a). In contrast, *C. jejuni* and *N. meningitidis* recombine at a much higher

rate, and in these species the identified clonal complexes vary considerably when the thresholds used to define them are changed. More detailed analyses of *Salmonella* have revealed that it is not clonal as was previously thought (J. M. Smith et al. 1993). From MLST data the rate of recombination to mutation (r/m) was estimated to be 30.2, suggesting very high recombination rates (Vos and Didelot 2009). However, a more recent analysis based on sequencing 10% of the core genome revealed lower r/m rates ranging from 2.95-0.15 depending on the lineage (Didelot et al. 2011). These estimates show that it is not easy to estimate the clonality of some species, and that there can be substantial within-species variation in recombination rates.

The increased resolution from whole-genome sequencing has enabled far more detailed analyses into bacterial population structure than was possible with previous methods. For predominantly clonal species such as *S. aureus*, phylogenetic analysis of of a diverse collection of isolates revealed that the species is composed of several distantly related clonal complexes, in agreement with MLST data (Figure 1.3a-b).

However, the power of whole-genome sequencing was illustrated by a landmark study on *S. aureus* ST239 (Harris et al. 2010), which provided a detailed analysis of a single *S. aureus* clonal complex (Figure 1.3c). The sample consisted of 63 isolates; 43 from a global collection recovered from 1982-2003, and 20 from Sappasithiprasong hospital in Thailand. 4310 single nucleotide polymorphisms (SNPs) were identified in the core genome of the 63 isolates, revealing considerable variation among isolates which were very closely related. Phylogenetic analysis showed that this variation was sufficient to identify transmission events on both a large scale (between continents) and a much finer scale (between patients within the same hospital) (Harris et al. 2010) (Figure 1.3c). This study was the first to show that whole-genome sequencing could be used to monitor outbreaks and identify transmission events; this has now become an essential tool for infectious disease control (Chan et al. 2012). It is worth noting that the isolates in this study are grouped into a single sequence type by MLST, and this serves as a dramatic example of the increased resolution offered by whole-genome sequencing over previous typing methods (Figure 1.3a-c).

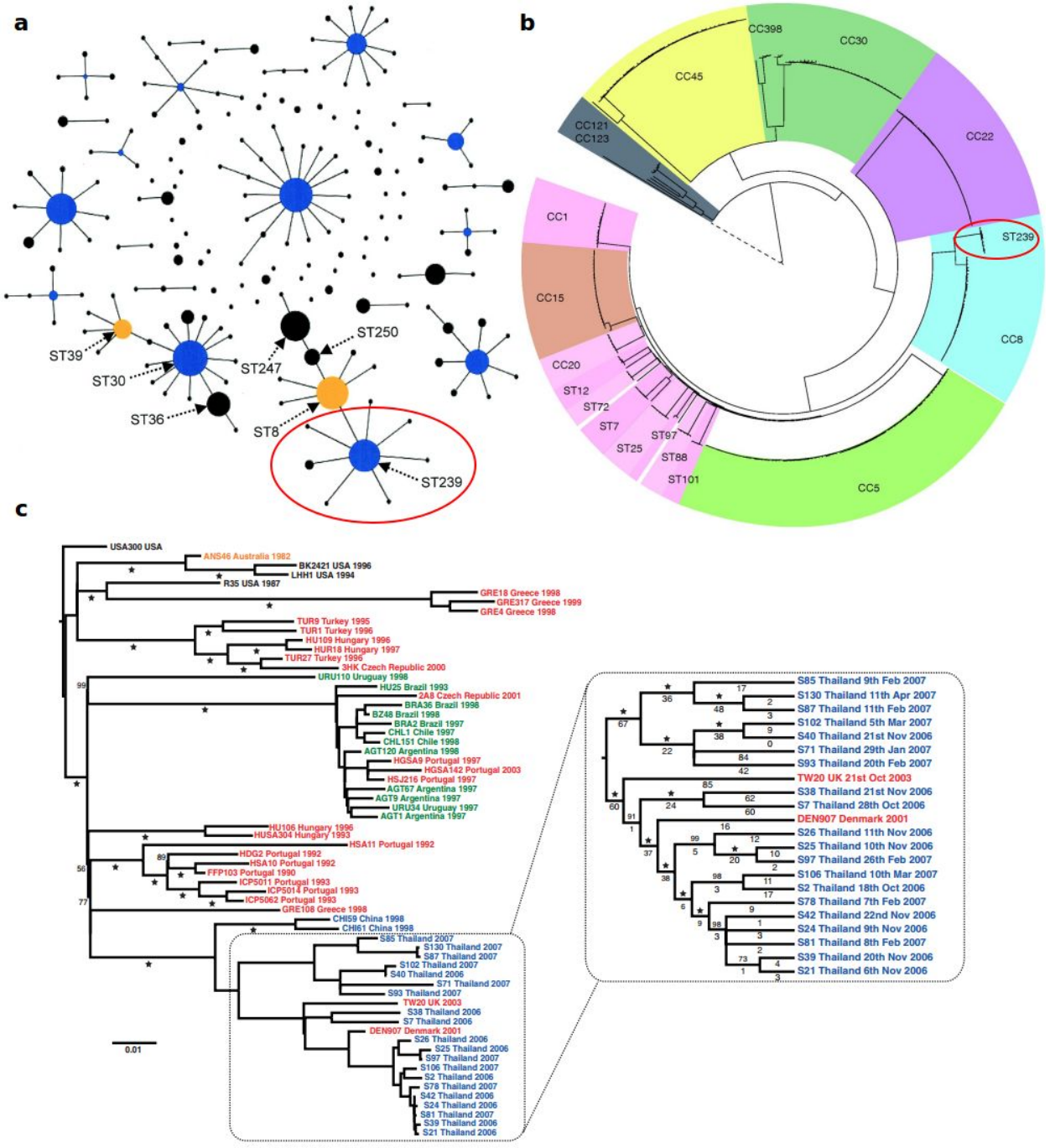

Figure 1.3: Bacterial population structure at different scales using different technologies. *S. aureus* is used to illustrate common features of bacterial populations. **a.** An eBURST diagram constructed from MLST data from a diverse collection of isolates. The circles are scaled according to the number of isolates. **b.** A phylogenetic tree constructed from whole-genome sequence data from a diverse collection of isolates. In both **a.** and **b.** ST239 is highlighted. **c.** A phylogenetic tree constructed from whole-genome sequence data from a collection of ST239 isolates. The detail present in this tree highlights the increased resolution of whole-genome sequence data compared with the MLST data shown in **a.** Figure adapted from (Feil et al. 2004; Aanensen et al. 2016; Harris et al. 2010).

Bacteria have large pan-genomes

In addition to extensive SNP variation among closely related strains, whole-genome sequencing has revealed large gene content variation between strains. This variation is known as the core genome (genomic loci present in all or nearly all strains), the accessory genome (genomic loci present in some strains), and the pan-genome (the core + accessory). Some species, such as *E. coli* and *K. pneumoniae* have large pan-genomes with many accessory genes. A recent study of 228 *E. coli* ST131 isolates revealed a pan-genome consisting of 11,401 genes, of which only 2,722 were core (present in all isolates) (McNally et al. 2016) (Figure 1.4c). Given that each isolate has approximately 4100 genes, only 65% of genes within a single isolate will be present in all other members of the collection. This level gene content variation is remarkable given that these isolates are all clustered into a single sequence type (ST131) by MLST. In contrast, another analysis of a single MLST type (*S. aureus* ST22), revealed that 86% of genes were core in a dataset of 193 isolates (Holden et al. 2013).

Compared to the previous analyses which focused on individual lineages, Holt *et al.* analysed a collection of 328 *K. pneumoniae* isolates which encompassed the variation within the species (Holt et al. 2015). This revealed a pan-genome of 29,886 genes, of which 1,888 were core (present in ≥95% of isolates) (Figure 1.4a). This means that less than half of the genes in an individual isolate are present in 95% of isolates (a typical *K. pneumoniae* isolate has 4500-5000 genes). Gene accumulation curves showed that the pan-genome was 'open', meaning that the estimate of the total number of genes will likely increase as more isolates are sampled and sequenced (Figure 1.4b). A striking feature of the *K. pneumoniae* pan-genome was that many accessory genes are likely to have been transferred from other species, notably the *Acinetobacter* and *Vibrio* genera (Holt et al. 2015).

It is not immediately clear why bacteria should require access to such a large gene pool, but environmental species such as *K. pneumoniae* occupy many different niches with varying conditions. McInerney *et al.* argue that large pan-genomes are the result of bacteria acquiring new genes, which give them the ability to migrate to new niches (McInerney, McNally, and O'Connell 2017). In a dataset of 228 *E. coli* ST131 isolates, accessory genes encoding metabolic functions were more common than those contained on selfish elements. This is consistent with accessory genes being advantageous for the bacteria, rather than selfish propagation of certain genomic elements (McInerney, McNally, and O'Connell 2017). In

contrast, Vos *et al.* showed that pan-genome size correlates with effective population size, suggesting that many accessory genes may be neutral (Andreani, Hesse, and Vos 2017).

Studies of pan-genomes are not strictly limited to protein-coding regions, but may also incorporate intergenic regions. In a landmark study by Oren *et al.*, adjacent genes were shown to be regulated by alternative intergenic alleles in different strains of *E. coli* (Oren et al. 2014). These alternative intergenic alleles frequently shared little sequence homology with each other (< 42%), and were incongruent with the phylogeny of the species, suggesting that they are transferred by recombination (Oren et al. 2014) (Figure 1.4d). These alternative alleles were shown to contain binding sites for different regulators, and were associated with differential expression of their downstream genes. This process was termed horizontal regulatory transfer (HRT), and is likely to be an important source of phenotypic variation in bacteria (Oren et al. 2014).

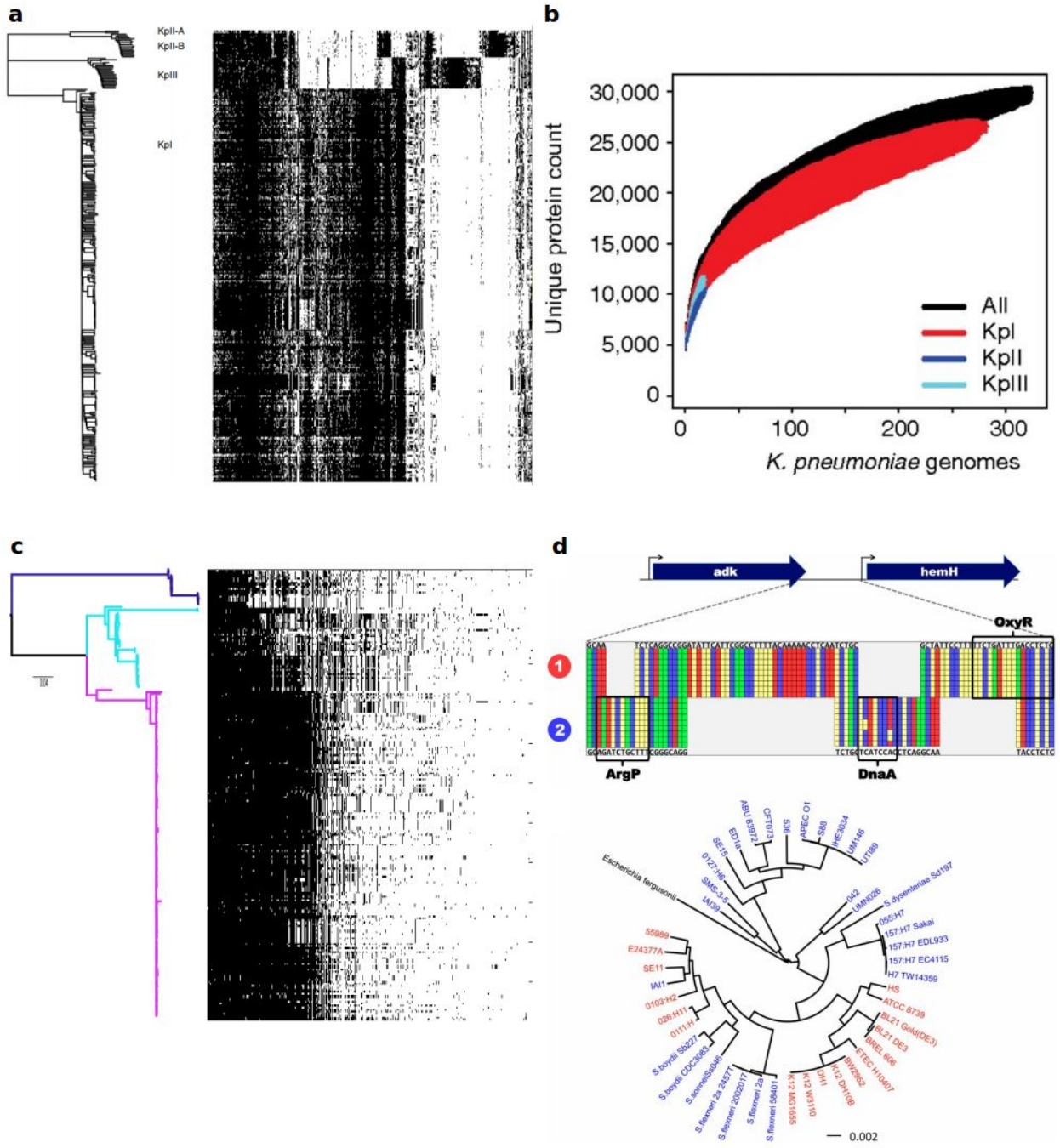

Figure 1.4: Bacterial pan-genomes. a. The pan-genome of *K. pneumoniae*. A phylogenetic tree is shown on the left, and each gene is shown by a black block on the right. The continuous block of genes on the left represent the core genes, and the others are accessory genes. **b.** Gene accumulation curves for *K. pneumoniae*. More genes are identified as more genomes are sequenced, indicating an 'open' pan-genome. **c.** The pan-genome of *E. coli* ST131. **d.** Horizontal regulatory switching in *E. coli*. Alternative intergenic regions are carried between two orthologous genes in different isolates, and these are incongruent with the tree. Figure adapted from (Holt et al. 2015; McNally et al. 2016; Oren et al. 2014).

Mechanisms of generating variation

Processes which generate variation in bacteria can be divided into two major groups, the first consisting of small variants of existing sequence and the second consisting of larger variants comprising the import/export of genomic elements, rearrangement of existing elements, or recombination between homologous sequences. Small variants can involve substitution of one nucleotide for another (for example an A -> T); these variants are known as single nucleotide polymorphisms (SNPs). Additionally, nucleotides can be inserted into or deleted from existing sequence (INDELs). These types of small variation are formed *de novo* by errors during the DNA replication process, and this distinguishes them from the larger variants (although SNPs or INDELs can also be introduced by recombination).

SNPs within genes can be divided into categories depending on the types of nucleotides exchanged, and how the mutation affects the encoded protein sequence. Transitions are mutations where the purine/pyrimidine status of the nucleotide doesn't change (e.g. A -> G or C -> T), whereas transversions are mutations between purines and pyrimidines (e.g. A -> T). Synonymous mutations are changes which do not affect the protein sequence due to the degenerate nature of the genetic code (for example CGA -> CGC both encode Arginine). Non-synonymous mutations are changes which do affect the protein sequence, for example CGA -> CCA results in an Arginine to Proline substitution. Nonsense mutations are changes where a premature stop codon is produced, for example CGA -> TGA.

There are 12 possible SNP types (each of the four nucleotides can change to one of three others), but these are not equally likely to occur. Two independent studies published simultaneously provided evidence that GC -> AT mutations are more frequent than the reverse in most bacterial species (Hershberg and Petrov 2010; Hildebrand, Meyer, and Eyre-Walker 2010) (Figure 1.5a). Hershberg *et al.* analysed large datasets corresponding to 5 clonal bacterial species, where many whole-genome sequences were available for each species. Importantly, within each species the isolates were very closely related, meaning that the mutations were recent, and thus suitable for measuring mutation biases. They found strong evidence that mutation was biased from GC -> AT in all species, and that this bias was primarily driven by C -> T (or reversibly G -> A) transitions (Hershberg and Petrov 2010). Hildebrand *et al.* analysed a much wider sample of 149 bacterial species, with the caveat that fewer isolates were included for each species, and the ages of the mutations were less certain. They found

that in all but the most AT rich bacterial species, mutation was biased from GC -> AT (Hildebrand, Meyer, and Eyre-Walker 2010). Both studies compared the observed GC contents of bacterial species with those predicted from the mutation biases, and found that observed GC contents were substantially higher than predicted by the mutation biases (Rocha and Feil 2010).

Small INDELs are formed by the addition or loss of one or more nucleotides into existing sequence. They are commonly formed by slippage of DNA polymerase during DNA replication, and therefore repeats are hotspots for INDEL mutations (Lin and Kussell 2012; Gu et al. 2010; Gragg, Harfe, and Jinks-Robertson 2002). Within mononucleotide repeats, INDEL formation rates increase exponentially with the length of the repeat tract, and the rate is higher in GC compared to AT mononucleotide repeats (Lin and Kussell 2012). GC repeats are thought to be more mutable than AT repeats because the mismatch repair system (MMR) is more able to remove slippage intermediates in AT repeat tracts compared to GC repeat tracts (Gragg, Harfe, and Jinks-Robertson 2002).

Horizontal gene transfer (HGT) is the process whereby genetic material is exchanged horizontally between cells, and is a significant contributor to bacterial evolution (Figure 1.5b-d). There are three principal mechanisms of HGT: transformation, conjugation, and transduction. Transformation is where exogenous DNA is taken into the cell and integrated into the genome by homologous recombination, a process which is facilitated by competence machinery encoded by the cell (Croucher et al. 2016) (Figure 1.5b). Double stranded DNA binds to an outer membrane protein, and is then translocated through a pore, during this process the complementary strand is degraded, and so the DNA transferred into the cytosol is single stranded. The single stranded DNA is then cleaved into fragments, which are bound to proteins to form a RecA nucleoprotein filament. This can then invade the duplex of the host chromosome, and homologous recombination can then occur (Johnston et al. 2014; Croucher et al. 2016). In general, bacteria tightly regulate transformation, although some species (e.g. *Helicobacter pylori*) are constitutively transformable (Johnston et al. 2014).

While transformation is driven by machinery encoded by the recipient cell, conjugation and transduction are driven by groups of genes from the donor cell which encode both the recombination machinery and other cargo genes; these groups of genes are transferred as units known as mobile genetic elements (MGEs) (Croucher et al. 2016; Juhas, Crook, and Hood

2008) (Figure 1.5c). In conjugative MGEs, the DNA is transferred through machinery consisting of a relaxase, a type IV secretion system (T4SS), and a type IV coupling protein (T4CP) (Guglielmini et al. 2011; Juhas, Crook, and Hood 2008). The relaxase nicks the DNA at the origin of transfer, this complex is then coupled to the T4SS by the T4CP, and the T4SS then translocates the DNA through the conjugative pilus into the cytoplasm of the recipient cell (Guglielmini et al. 2011; Juhas, Crook, and Hood 2008). DNA transferred through conjugation can either be autonomously replicating (such as a plasmid), or not (in which case the DNA is integrated into the recipient chromosome). Whereas transformation mostly transfers small fragments of DNA, conjugation transfers much larger fragments. For example, in *Streptococcus agalactiae*, DNA fragments of up to 334 Kb are transferred by conjugation (Brochet et al. 2008).

Transduction involves the transfer of DNA through an MGE encoded bacteria-infecting virus particle (phage), which transfers the DNA by injecting it into the recipient bacterial cell (Croucher et al. 2016; Feiner et al. 2015) (Figure 1.5d). It has been estimated that there are 10^{31} phage in the biosphere, making them the most abundant known biological entities (Rohwer and Edwards 2002). In order to defend themselves against this threat, bacteria have developed an immune system consisting of clustered regularly interspaced short palindromic repeats (CRISPRs). When a bacterium is infected by a phage, some of the phage DNA is stored between the CRISPR repeats, and this renders the bacterium resistant to subsequent infection by the same phage (Vale and Little 2010). This arms race between bacteria and phage has had a profound impact on bacterial genomes, and as a result many genes within a bacterial genome are phage-derived. In *E. coli* O157 strain Sakai, up to 16% of the chromosomal DNA is phage-derived (Canchaya et al. 2003). Phage-derived genes are often virulence factors, such as the shiga toxin in enterohaemorrhagic *E. coli* (Canchaya et al. 2003). Other examples of important phage-derived MGEs include the *S. aureus* pathogenicity islands (SaPIs), which encode the toxic shock protein and enterotoxin B (Penadés et al. 2015).

HGT can influence the diversification of bacteria by transferring novel genes or variant alleles of existing genes. The transfer of novel genes results in substantial differences in gene content between isolates, resulting in large pan-genomes. In contrast, the transfer of variant alleles influences the diversification of homologous sequences which are shared within a species. The relative rates of homologous recombination and *de novo* point mutation determine which is the major driver of bacterial diversification, and this quantity is often expressed as r/m (the relative

number of mutations introduced by recombination compared to mutation). A study by Feil *et al.* showed that diversification in *S. aureus* is driven primarily by mutation, whereas in *S. pneumoniae* it was driven primarily by recombination (Feil et al. 2003). Vos *et al.* used MLST data to estimate r/m in a large number of bacterial species and found extensive variation in r/m rates (Vos and Didelot 2009). Some species were extremely recombinogenic, such as *Vibrio parahaemolyticus* (r/m = 39.8) and *S. pneumoniae* (r/m = 23.1). Some species had intermediate rates, such as *Campylobacter jejuni* (r/m = 2.2) and *Enterococcus faecium* (r/m = 1.1), and some species had very low recombination rates, such as *Staphylococcus aureus* (r/m = 0.1) and *Clostridium difficile* (r/m = 0.2) (Vos and Didelot 2009).

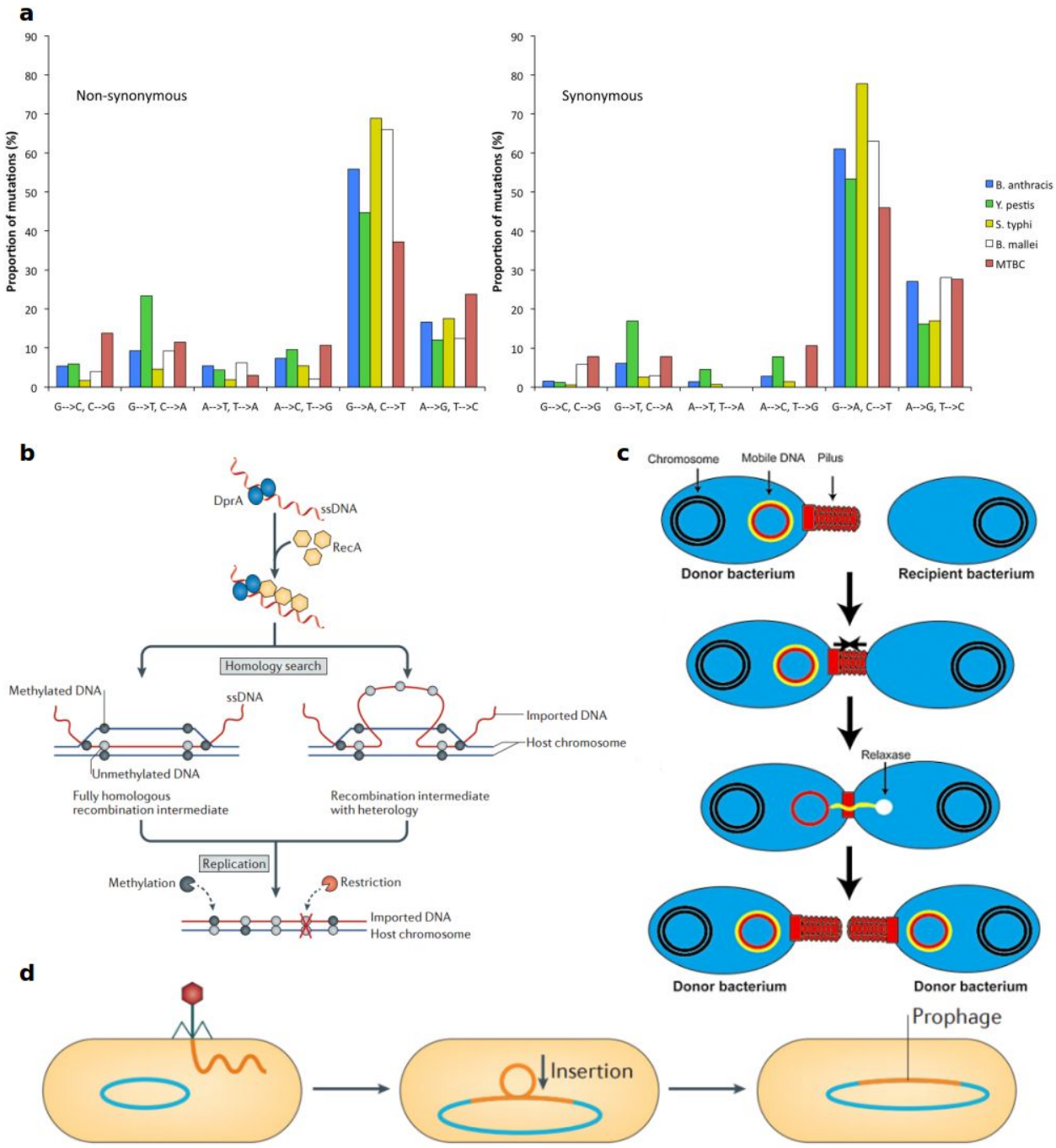

Figure 1.5: Mechanisms of generating variation in bacteria. a. Rates and biases of single nucleotide polymorphisms in five different species. Non-synonymous mutations are shown on the left, and synonymous mutations are shown on the right. **b.** The mechanism of transformation. **c.** The mechanism of conjugation. **d.** The mechanism of transduction. Figure adapted from (Hershberg and Petrov 2010; Johnston et al. 2014; Juhas, Crook, and Hood 2008; Feiner et al. 2015).

Population genetics

Population genetics is the study of allele frequency change, and describes the interactions between four major elements: mutation, HGT, drift, and selection. Mutation is the *de novo* occurrence of genetic variation (SNPs, INDELs). HGT is the import/export of genetic variation into or out of the population, or the reassortment of existing variation within the population. Drift is the change in allele frequencies caused by random sampling of the population. Selection is the understanding that genetic variants may have fitness effects, and these effects influence the change in allele frequency of that variant (Casillas and Barbadilla 2017; Charlesworth 2009). Briefly, mutation and recombination provide the raw input for evolution, and drift and selection influence how that variation is assorted (the output).

How this assortment happens depends on two factors, the selective coefficient (s), and the effective population size (N_e). The selective coefficient is relative difference in fitness between two variants; a variant with a fitness advantage of 1% is represented as $s = 0.01$. The effective population size refers to the number of genotypes which contribute variation to future generations in an ideal population (random mating, constant population size), and has an important effect on the relative power of selection and drift in determining the fate of a variant (Casillas and Barbadilla 2017; Charlesworth 2009). The power of drift is determined by the law of large numbers, and so is dependent on N_e. If $s < 1/N_e$ (i.e. when the selective coefficient is less than 1 divided by the effective population size), then drift can overpower selection, and alternatively if $s > 1/N_e$, then selection will be effective (Casillas and Barbadilla 2017; Charlesworth 2009). Thus, whether a variant is selected or not depends on its fitness effect and the size of the population. A variant of $s = 0.01$ will be selected if the population contains > 100 individuals, whereas a much more subtle variant of $s = 0.000001$ will only be selected if the population contains > 1,000,000 individuals, otherwise its fate will be governed by drift. This means that selection is a more powerful force in large populations, and more subtle genetic variants can be selected.

Measuring selection

One way to measure selection is to use *a priori* assumptions about mutation types in order to construct tests. dN/dS is one of the most well used tests for measuring selection, and compares the rates of observed non-synonymous (dN) and synonymous (dS) mutations (Casillas and Barbadilla 2017; Hurst 2002). It is based on the assumption that synonymous mutations (which

do not result in an amino acid change) are neutral (or at least more neutral than non-synonymous mutations). Thus, in the absence of selection dN/dS = 1, as the per sites rates of the two mutation types are equal. However, if dN/dS < 1, then this suggests that the observed rate of non-synonymous mutation is lower than that of synonymous mutation, and this is interpreted as purifying selection against deleterious non-synonymous mutations, removing them from the population. Alternatively, if dN/dS > 1, this is interpreted as positive selection for advantageous non-synonymous mutations (Casillas and Barbadilla 2017; Hurst 2002).

However, dN/dS is not a useful statistic if part of a protein is under positive selection (for example an antigenic domain on the cell surface), and another part is under purifying selection (for example a domain which anchors the protein into the membrane). In this scenario the purifying and positive selection would cancel out, possibly resulting in dN/dS = 1, and this could then be falsely interpreted as no selection on the protein (Hurst 2002). The principle of this test is not limited to only non-synonymous and synonymous mutations; variants of the test have used non-coding and INDEL mutations in place of non-synonymous mutations (Zhou et al. 2014; Feng and Chiu 2014).

Another way to measure the effect of a variant is to statistically associate it with a particular phenotype in a genome-wide association study (GWAS). In a GWAS study, variants are tested for association against phenotypes, and for an association to be valid it must remain significant after correction for both the population structure and multiple testing. Although these associations are not strictly evidence of selection, often phenotypes known to be strongly selected for are identified (for example antibiotic resistance), and so in these cases it is reasonable to assume that these variants are selected (Lees et al. 2016; Earle et al. 2016). GWAS has been successfully applied to in human genetics for many years (Stranger, Stahl, and Raj 2011), but in bacteria the problem of correcting for population structure is more difficult, and GWAS studies have only recently become possible (Sheppard et al. 2013). In bacteria, binary fission and limited recombination result in a strong clonal frame where many variants are inherited together. This makes it difficult to distinguish between causal and co-inherited variants, as many co-inherited variants may be associated with a particular phenotype. Several methods have recently been proposed to correct for this problem, (Sheppard et al. 2013) and treeWAS (Collins and Didelot 2017) use simulated phylogenetic trees, bugwas (Earle et al. 2016) uses principal components analysis (PCA) to detect and correct for lineage effects, and SEER (Lees

et al. 2016) uses PCA based on Kmers. Scoary (Brynildsrud et al. 2016) tests the association of variably present genes within the pan-genome against phenotypes.

Identifying important variants through GWAS studies requires that phenotypes have been measured. However, it is possible to identify epistatic interactions (defined as non-additive interactions between variants) without first measuring phenotypes. As with GWAS studies, population structure control must be applied to ensure that co-occurring variants are not simply the result of linkage on the genome. These methods were first applied to *V. parahaemolyticus*, where strong epistatic interactions between a type VI secretion system and biofilm forming genes were found (Cui et al. 2015). More recently, a method based on direct coupling analysis (genomeDCA) was developed, and this identified interactions between the penicillin binding proteins in *S. pneumoniae* (Skwark et al. 2017).

Selection shapes bacterial genomes

Owing to their small sizes, ubiquity, and short generation times, many bacterial species have large long term effective population sizes; for example in *E. coli* N_e has been estimated to be 25,000,000 (McInerney, McNally, and O'Connell 2017; Charlesworth 2009). In populations of this size, selection is powerful enough to influence very subtle genetic variants, such as alternative codons which encode the same amino acid. This is shown by an analysis of 80 bacterial species, where the strength of selected codon bias in *E. coli* is among the highest of all species tested (Sharp et al. 2005).

In an analysis of core genes from 6 bacterial species, Rocha *et al.* showed that (with the exception of the low diversity *Chlamydia pneumoniae* and *Mycobacterial* species), dN/dS values were 0.1 or lower (Rocha et al. 2006). This is convincing evidence of strong purifying selection acting on these species. More strikingly, dN/dS decreased with divergence time, such that comparisons between distantly related isolates had lower dN/dS values than those between more closely related isolates. This is shown as increasing dS/dN values in Figure 1.6a. This observation was also recovered from both deterministic and stochastic models, where values of s for non-synonymous mutations, and N_e were varied across a wide range of parameters (Rocha et al. 2006). The concordance between the observed and simulated data, and the effect of dN/dS decreasing with time, are consistent with expectations from the nearly neutral theory of evolution (Ohta 1973). If most non-synonymous mutations are slightly deleterious (i.e. have

small negative values of *s*), then they will not be removed from the population immediately, but will persist for some time. Between closely related bacterial isolates (such as those from the same clonal complex), many of the mutations are recent, and selection has not yet had time to purge them from the population, and so they are observed. A clear example of this is found in *S. aureus*, where comparisons between isolates from the same clonal complex have dN/dS values of approximately 0.5, whereas those from different clonal complexes have values < 0.1 (Castillo-Ramírez et al. 2011).

In addition to time, recombination has been shown to affect inferences of dN/dS (Castillo-Ramírez et al. 2011). In *S. aureus* ST239, recombination is infrequent within the core genome, but more prevalent within the accessory genome. In the core genome, high dN/dS values (of approximately 0.7) were observed, consistent with these isolates being very recently diverged. In contrast, dN/dS values in accessory regions of the genome were much lower (0.2) (Figure 1.6b). Additionally, within the core genome, non-recombined regions had higher dN/dS values than regions which had been subjected to homologous recombination from distant lineages (Castillo-Ramírez et al. 2011). These results are consistent with the model of ongoing purifying selection, as SNPs within recombined regions are likely to be older, and therefore have been subject to selection for longer than those which have emerged *de novo* only recently.

In bacterial species where the genome consists of multiple chromosomes, there is commonly one larger, more conserved chromosome, and other smaller chromosomes. Cooper *et al.* investigated the possibility that different chromosomes are subject to varying levels of selection (Cooper et al. 2010). Core genes from *Burkholderia* (three chromosomes) and *Vibrio* (two chromosomes) were analysed. It was found that evolutionary rates (dN and dS) were faster on the secondary chromosome compared to the primary chromosome (Figure 1.6c-d). Codon bias was also stronger on primary chromosomes than on secondary chromosomes. By comparison of orthologs from species with only a single chromosome, it was found that the genes on secondary chromosomes are inherently faster evolving than those on primary chromosomes (Cooper et al. 2010). This suggests that more conserved genes are localised to the primary chromosome, and those which are less conserved are located to the secondary chromosome. Together, these results show that the strength of purifying selection can vary between chromosomes from the same species, and have implications for genome organisation in bacteria.

Because the strength of selection depends on *s* and N_e, changes in ecological niches are likely to influence selection, for example if N_e decreases then selection will become weaker. This was investigated by Balbi *et al.*, who used *E. coli* and *Shigella* as a model for ecological shift (Balbi, Rocha, and Feil 2009). *E. coli* is a ubiquitous, diverse species of enteric bacterium, and *Shigella* is the name given to lineages of *E. coli* which have acquired the pINV plasmid and adopted an intracellular lifestyle. This shift in lifestyle could have reduced the N_e of *Shigella* species, or it could have rendered some genes unimportant, thus reducing *s* in those genes, or a combination of both. In both cases selection would be weaker. Additionally, the shift may present greater possibility for adaptive evolution in response to the new environment. It was found that the *Shigella* isolates were carrying more deleterious mutations than the *E. coli* isolates, consistent with relaxed or inefficient purifying selection (Balbi, Rocha, and Feil 2009). The *Shigella* isolates had higher dN/dS values, more GC -> AT mutations (consistent with mutation bias), and more transitions than their *E. coli* counterparts, and consistent with a model of ongoing purifying selection, these quantities all decreased with divergence time (Balbi, Rocha, and Feil 2009) (Figure 1.6e-f).

There are several studies which show that bacterial GC contents are typically higher than the equilibrium GC content calculated from mutation biases (given the GC -> AT mutation bias in bacteria) (Hershberg and Petrov 2010; Hildebrand, Meyer, and Eyre-Walker 2010; Rocha and Feil 2010; Balbi, Rocha, and Feil 2009). It has been proposed that this discrepancy is the result of selection for higher GC contents, maintaining the GC content above the equilibrium state. In *E. coli* and *Shigella*, GC -> AT mutations are more common than the reverse, but this effect is less pronounced with increasing divergence time, suggesting that deleterious AT increasing mutations are removed by selection over time (Balbi, Rocha, and Feil 2009). Additionally, the number of AT increasing mutations is higher in *Shigella* than *E. coli*, which is consistent with other features of reduced selection in *Shigella* (high dN/dS values) (Balbi, Rocha, and Feil 2009). The selective advantage of high GC contents in bacteria is not clear (Rocha and Feil 2010). One study tested the effect of GC content on gene expression in *E. coli* (Raghavan, Kelkar, and Ochman 2012). Multiple copies of a gene were synthesised to differ in GC content, without affecting codon bias, and the expression of these variants was measured. The expression level was dependent on GC content, with the high GC variants being the most highly

expressed (Raghavan, Kelkar, and Ochman 2012). This gives a glimpse into the possible selective advantage of high GC content, but many questions remain.

More extreme examples of genomic changes come from bacterial endosymbionts, which have reached a ‘point of no return’ and are committed to coevolution with their hosts (Balbi, Rocha, and Feil 2009). This can manifest in extreme genome reduction, for example a 90% reduction in genome size in *Buchnera aphidicola* (McCutcheon and Moran 2011; Moran and Mira 2001). Endosymbiont genomes are typically AT rich, for example the GC content of *Carsonella ruddii* is only 16.5%, and these genomes are typically closer to mutational equilibrium than their free-living counterparts (Nakabachi et al. 2006).

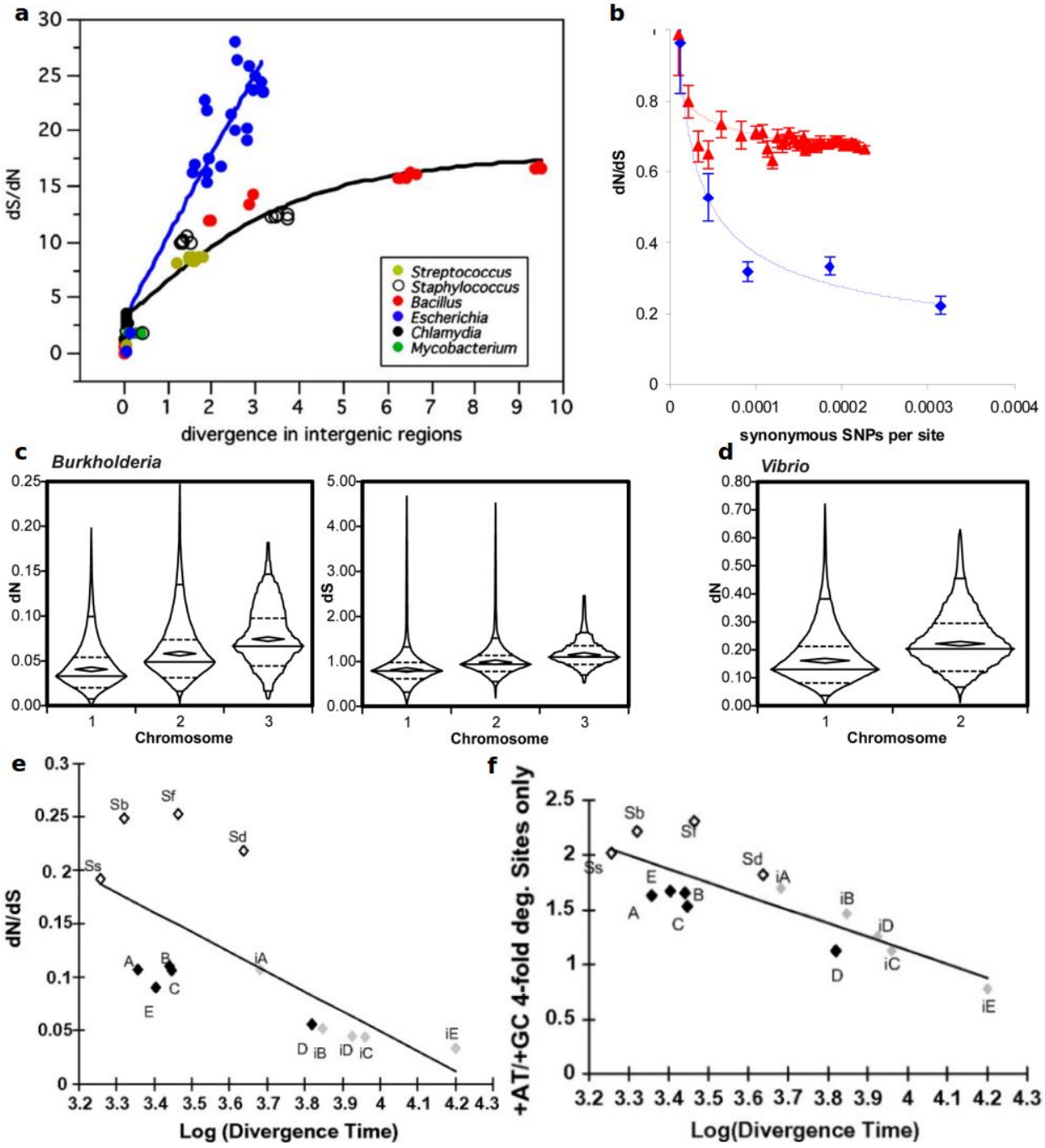

Figure 1.6: Selection shapes bacterial genomes. a. Increasing dS/dN (decreasing dN/dS) values in six bacterial species. dS/dN was plotted against divergence in intergenic regions as a proxy for divergence time. **b.** dN/dS values in the core (red) are higher than those in the accessory (blue) part of the genome in *S. aureus* ST239. **c.** Lower evolutionary rates on chromosome 1 compared with the secondary chromosomes in *Burkholderia* and **d.** *Vibrio* species. **e.** Higher dN/dS values in *Shigella* genomes (white diamonds) compared with *E. coli* (black diamonds) or internal branches (grey diamonds). dN/dS decreases with divergence time. **f.** Same as for **e.**, but for +AT/+GC mutations instead of dN/dS. Figure adapted from (Rocha et al. 2006; Castillo-Ramírez et al. 2011; Cooper et al. 2010; Balbi, Rocha, and Feil 2009).

The work presented in this thesis

The review of current literature reveals that bacterial genomes evolve under strong selective constraints; they are compact, and selection is measurable on features with very subtle fitness effects (codon bias and AT skew). They are also diverse, with gene content varying widely between individuals from the same species. The work presented in this thesis has attempted to advance our understanding of bacterial genome evolution by investigating these two themes in greater detail. This has predominantly involved analysing the non protein-coding or 'intergenic' component of the genome. In chapter 1, intergenic regions (IGRs) were analysed both *en masse*, and subdivided into individual regulatory elements, in order to test for signals of selection in the core genomes of a range of bacterial species. In chapter 2, the intergenic component of the pan-genome was considered, and a new tool, Piggy, was developed to facilitate these analyses. In this chapter RNA-seq data was combined with genomic data to investigate the association between changes in IGRs and gene expression. Chapter 3 focuses on the effect of selection on introgressed DNA in *Helicobacter pylori*, and chapter 4 presents a detailed analysis of compensatory evolution in terminator sequences.

Chapter 2

Comparative analyses of selection operating on non-translated intergenic regions of diverse bacterial species

The work presented in this chapter is published as a peer-reviewed publication at:

Thorpe, Harry A., Sion C. Bayliss, Laurence D. Hurst, and Edward J. Feil. 2017. "Comparative Analyses of Selection Operating on Non-Translated Intergenic Regions of Diverse Bacterial Species." ***Genetics*****, March. doi:10.1534/genetics.116.195784.**

Commentary text

The work in this chapter provides a comprehensive analysis of both purifying and positive selection on intergenic sites within bacterial genomes. 6 diverse bacterial species were analysed in order to enable comparisons between species to be made. Intergenic sites were analysed both *en masse*, and were also divided into different regulatory elements in order to compare differing selective pressures on these elements. Widespread purifying selection was found on intergenic sites, the strength of which varied according to the class of intergenic site. The statement of authorship for this chapter can be found in the Appendix, supplementary form SF1.

Abstract

Non-translated intergenic regions (IGRs) comprise 10-15% of bacterial genomes, and contain many regulatory elements with key functions. Despite this, there are few systematic studies on the strength and direction of selection operating on IGRs in bacteria using whole-genome sequence datasets. Here we exploit representative whole-genome datasets from six diverse bacterial species; *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Mycobacterium tuberculosis*, *Salmonella enterica*, *Klebsiella pneumoniae* and *Escherichia coli*. We compare patterns of selection operating on IGRs using two independent methods; the proportion of singleton mutations, and the dI/dS ratio; where dI is the number of intergenic SNPs per intergenic site. We find that the strength of purifying selection operating over all intergenic sites is consistently intermediate between that operating on synonymous and non-synonymous sites. Ribosome binding sites and non-coding RNAs tend to be under stronger selective constraint than promoters and rho-independent terminators. Strikingly, a clear signal of purifying selection remains even when all these major categories of regulatory elements are excluded, and this constraint is highest immediately upstream of genes. Whilst a paucity of variation means that the data for *M. tuberculosis* are more equivocal than for the other species, we find strong evidence for positive selection within promoters of this species. This points to a key adaptive role for regulatory changes in this important pathogen. Our study underlines the feasibility and utility of gauging the selective forces operating on bacterial IGRs from whole-genome sequence data, and suggests that our current understanding of the functionality of these sequences is far from complete.

Introduction

The ability to generate whole-genome sequence datasets from very large samples of bacterial isolates recovered from natural populations provides unprecedented power to dissect evolutionary processes. Although tests for selection are routinely carried out on the ~85-90% of bacterial genomes corresponding to protein-coding sequences, attempts to measure the strength and direction of selection operating on non-translated intergenic regions (IGRs) are far less common. Notable exceptions include the study by Molina *et al.*, who demonstrated that the number of regulatory elements per IGR is independent of genome size within bacteria, but also noted a surprising level of purifying selection operating on bacterial IGRs (Molina and Van Nimwegen 2008). However, as this study pre-dates the advent of next-generation sequencing, very large whole-genome datasets for single species were unavailable at that time. More

recently, Luo *et al.* found evidence for purifying selection within IGRs of a small sample (n=13) of group A streptococcal genomes (Luo et al. 2011) and Degnan *et al.* found strong evidence for sequence conservation within IGRs of eight Buchnera genomes (Degnan, Ochman, and Moran 2011). This observation is particularly striking in Buchnera, as it is an endosymbiont and so likely has a small effective population size (N_e). Together, these studies challenged the view held for many years that intergenic sites provide a valid proxy for neutrality (Wang and Chen 2013; Hu, Lan, and Reeves 2006; S. Fu et al. 2015).

Whole-genome sequence datasets for bacteria now routinely encompass many hundreds of genomes for a single species, although currently these data have remained almost completely untapped with respect to examining selection on IGRs. In fact, despite the studies mentioned above, many commonly used pipelines and databases by default exclude IGRs altogether, with the focus instead on defining sets of 'core' or 'accessory' genes (CDSs), upon which phylogenetic, epidemiological or evolutionary analyses are then carried out (M. C. J. Maiden et al. 2013; Jolley and Maiden 2010; Sheppard, Jolley, and Maiden 2012; Feil 2015; M. C. J. Maiden and Harrison 2016; Page et al. 2015). One explanation for this apparently casual dismissal of IGRs is a lack of 'off-the-shelf' methodology to measure selection on these sequences; the standard approach for protein-coding sequences, the dN/dS ratio, being invalid. In addition, there may be a prevailing sense that IGRs are technically challenging to work with, owing to low levels of constraint, poor annotation and a high frequency of indels. The approach by Fu *et al.* is a rare exception which challenges this view. These authors generated a core genome consisting of both genes and IGRs for *Salmonella enterica* serovar *Typhimurium*, and in so doing demonstrated the feasibility of incorporating IGRs into routine analysis. Furthermore, these authors showed that IGRs contribute meaningful signal to increase discriminatory power for phylogenetic and epidemiological analyses (S. Fu et al. 2015).

The paucity of studies aimed at systematically measuring selection on non-translated IGRs is strikingly at odds with the many recent examples demonstrating the phenotypic impact of mutations in riboswitches, small RNAs, promoters, terminators, and regulator binding sites (Waters and Storz 2009). Single nucleotide polymorphisms (SNPs) or small insertion/deletions (INDELs) within these elements can have major phenotypic consequences. For example, in a recent GWAS study, 13 intergenic SNPs were found to be significantly associated with toxicity in Methicillin-resistant *Staphylococcus aureus* (MRSA), and four of these were experimentally

validated (Laabei et al. 2014). In *Mycobacterium tuberculosis*, mutations within the *eis* promoter region increase expression of Eis, an enzyme which confers resistance to kanamycin and promotes intracellular survival (Casali et al. 2012). In addition to those studies focussing on naturally occurring mutations, knock-out experiments on regulatory RNAs have also confirmed their key roles in virulence and other important phenotypes such as competence. For example, the *Salmonella* sRNA IsrM is important for invasion of epithelial cells and replication inside macrophages (Gong et al. 2011). In *S. aureus*, the Sigma B-dependent *RsaA* sRNA represses the global regulator *MgrA*; this decreases the severity of acute infection and promotes chronic infection (Romilly et al. 2014). In *S. pneumoniae*, the *srn206* non-coding RNA is involved in competence modulation (Acebo et al. 2012).

These well characterised regulatory elements are clearly expected to be under strong purifying selection, but there remain no broad measures of the commonality of constraint operating on IGRs at an intra-species level. There is also currently little understanding of which intergenic regulatory elements are under strongest selection, whether a signal of selection can be detected even for those intergenic sites for which there is no known function, or to what extent positive (as well as negative) selection may be operating on IGRs. Here we use two independent approaches to address these questions. The first method is based on the established logic of site frequency spectra (the Proportion of Singleton Mutations; PSM), whilst the second is a modification of dN/dS (dI/dS; where dI is the number of intergenic SNPs per intergenic site). We apply these approaches to large whole-genome datasets from six diverse bacterial species; *Escherichia coli*, *Staphylococcus aureus, Salmonella enterica*, *Streptococcus pneumoniae*, *Klebsiella pneumoniae*, and *Mycobacterium tuberculosis*. With the exception of *M. tuberculosis*, our results demonstrate that the overall strength of selective constraint on intergenic sites in bacteria is intermediate between that operating on synonymous and non-synonymous sites. This observation does not significantly alter even when all major regulatory IGR elements are removed from the analysis, consistent with a substantial level of cryptic functionality in these sequences. We also compare the strength and direction of selection operating on different types of regulatory element within IGRs, and note strong evidence of positive selection acting on promoters in *M. tuberculosis*.

Methods

Data selection

For *S. aureus, S. pneumoniae, K. pneumoniae,* and *M. tuberculosis*, isolates were selected from recently published data (Reuter et al. 2015; Chewapreecha et al. 2014; Holt et al. 2015; Casali et al. 2014). For *S. enterica*, isolates were selected from those whole-genome sequenced routinely by the Gastrointestinal Bacteria Reference Unit at Public Health England. Recent large-scale bacterial genome sequencing projects have been primarily motivated by efforts to understand features which are important for public health, such as disease transmission, virulence, and antibiotic resistance. Consequently, the datasets may be poorly representative of the population as a whole, with disproportionate weight given to lineages of particular clinical significance. For example the *S. aureus* data were generated as part of a retrospective study of hospital-acquired methicillin resistant *S. aureus* MRSA in the UK (Reuter et al. 2015), and the majority of these isolates corresponded to a single clonal lineage, CC22 (EMRSA-15). We therefore subsampled the datasets to include each major lineage and a random sample from the over-represented clonal complexes. A complete list of all isolates used in the analysis is given in Table S2.1.

Sequencing, mapping and SNP calling

For each species except *E. coli*, reads were downloaded from the ENA (http://www.ebi.ac.uk/ena). For *E. coli*, completed genome sequences were downloaded from NCBI, and sheared into reads with ArtificialFastqGenerator (Frampton and Houlston 2012). The isolates were mapped against a single reference genome for each species (as shown in Table 2.1) using SMALT-0.7.6 (https://sourceforge.net/projects/smalt). SAMtools-0.1.19 (Li et al. 2009) was used to produce Variant Call Format (VCF) files, which were filtered to call SNPs. SNPs were only called if they passed all of the following thresholds: depth >= 4, depth per strand >= 2, proportion of reads supporting the SNP >= 0.75, base quality >= 50, map quality >= 30, af1 >= 0.95, strand bias >= 0.001, map bias >= 0.001, tail bias >= 0.001. Consensus Fasta sequences were then used to produce an alignment for each species. [N.B. The mapping was performed in collaboration with Sion Bayliss, University of Bath, Bath, UK.]

Validation of singleton SNPs

As singleton SNPs are potentially vulnerable to poor quality data, we performed a thorough analysis of the SNPs to validate their quality. This was based on analysing three metrics: depth

of coverage, proportion of reads supporting the variant, and the Phred Quality (Q) score in both singletons and non-singletons. Q is related to the per-base error probability according to the following equations:

$Q = -10log_{10}P$ and $P = 10^{\frac{-Q}{10}}$

In illumina reads, the per-base Q score is approximately Q30, equating to one error every 10^{-3} bases. However, this error rate is substantially reduced by sequencing to high coverage, and then mapping the reads to the reference genome. For each species in our analysis (with the exception of *E. coli*), the sequencing depth was 50-100x per isolate. For *E. coli*, we simulated reads with no errors, using the complete genome sequences, and then mapped these synthetic reads to the standard reference genome (MG1655).

To validate our SNPs, we used the mapping information in the Variant Call Format (VCF) files. We focused on three metrics, the depth of coverage, the proportion of reads supporting the variant, and the Q score for the position (which takes into account the per-base-per-read error rate, and the coverage at the position). We split our SNPs into singletons and non-singletons to check for singleton associated biases.

IGR identification and core genome definition

Each reference genome was annotated using Prokka-1.11 (Seemann 2014). This annotation was used to extract genes and IGRs (IGRs > 1000 bp in length were excluded), and three core sets of genes and IGRs were defined for each species. tRNA and rRNA genes were excluded from all analyses. The 'relaxed core' consisted of all genes and IGRs, the 'intermediate core' consisted of all genes and IGRs with > 90% sequence present in > 95% of isolates, and the 'strict core' consisted of genes and IGRs with > 90% sequence present in > 99% of isolates.

Calculation of dN/dS and dI/dS

Core gene and intergenic sequences were extracted from the alignments and concatenated to produce gene and intergenic alignments (reverse oriented genes were reverse complemented so all genes were in sense orientation). The codons within the gene alignment were shuffled, and the gene alignment was split into two (referred to as a and b). The YN00 program from the PAML suite (Ziheng Yang 2007) was used to calculate dN/dS values by the Nei and Gojobori (1986), and Yang and Nielson (2000) methods in a pairwise manner for both gene alignments a and b (Nei and Gojobori 1986; Z. Yang and Nielsen 2000). The results were almost identical,

and so we used the Nei and Gojobori (1986) method as it was computationally less demanding, and enabled the results to be compared directly with the dI values. SNPs were counted between isolates in a pairwise manner from the intergenic alignment, and dI was calculated by dividing the number of SNPs by the length of the alignment, before applying a Jukes-Cantor distance correction (Jukes and Cantor 1969). For both the gene and intergenic alignments Ns were removed from the alignment in a pairwise manner to ensure that all possible data was used. The dS values from gene alignment a were used to calculate dN/dS and dI/dS, and the dS values from gene alignment b were used as a proxy for divergence time. This ensured that when plotting dN/dS and dI/dS against dS, the dS values on each axis were calculated independently, thus controlling for statistical non-independence.

Correcting dN/dS and dI/dS calculations for mutation biases and base composition

In order to confirm that the model we used to calculate dN/dS and dI/dS accurately reflects the null expected under neutrality, we simulated neutral divergence of the reference genomes based on the observed mutational spectra, then recalculated dN/dS and dI/dS from the simulated sequences. Any deviation from parity (dN/dS = 1) reflects the fact that we have not accurately incorporated mutation bias, in particular the strong AT pressure in bacterial genomes (Balbi, Rocha, and Feil 2009; Hershberg and Petrov 2010; Hildebrand, Meyer, and Eyre-Walker 2010) and base composition into our models. However, by calculating the magnitude of the deviation between the simulated sequences and parity we can correct for this bias.

We first calculated the per-site mutation bias for the 6 mutation types for each species (Figure S2.1). We then simulated neutral mutations on a sequence of concatenated genes and IGRs to a divergence of 1% from the original sequence for 50 replicates. We then calculated dN/dS and dI/dS between pairwise comparisons of these 50 replicates. This gave us an expectation of dN/dS and dI/dS under neutral conditions, taking into account mutation biases and base composition. We then computed observed/expected (simulated) dN/dS and dI/dS ratios, thus providing corrected estimates. We did this for alignments of each intergenic element considered (promoters, terminators, ribosome binding sites, non-coding RNAs, and unannotated sites) to also correct these estimates.

Ribosome binding site, promoter, non-coding RNA, and terminator annotation

Promoter and terminator predictions were obtained using the PePPER webserver (de Jong et

al. 2012). Non-coding RNA annotations were obtained from the reference genome annotation GFF file produced by Prokka, where they were labelled as 'misc_RNA'. Ribosome binding site annotations were predicted using RBSfinder (Suzek et al. 2001).

Code availability and computation

All of the code used in the analysis is available at https://github.com/harry-thorpe/Intergenic_selection_paper under the GPLv3 license. The complete analysis can be reproduced by running a single script, using any alignment and annotation files as inputs. Full instructions are available in the GitHub repository. All computations were performed on the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) (Connor et al. 2016). All figures were produced using the R package ggplot2 (Wickham 2009).

Results

Species and data selection

We used existing large whole-genome sequence datasets for six diverse bacterial species: *Escherichia coli*, *Staphylococcus aureus, Salmonella enterica*, *Streptococcus pneumoniae*, *Klebsiella pneumoniae*, and *Mycobacterium tuberculosis*. These species are diverse in terms of phylogeny (representing Gram-positive and Gram-negative taxa), in terms of population structure (ranging from the highly clonal *M. tuberculosis* to the freely recombining *S. pneumoniae*), and in terms of ecology. The *K. pneumoniae* and *E. coli* data include isolates from environmental sources and disease, the *S. aureus* and *S. pneumoniae* data includes isolates from asymptomatic carriage, and *M. tuberculosis* is an intracellular pathogen. The GC content of these species range from 32.9% (*S. aureus*) to 65.6% (*M. tuberculosis*) (Table 2.1). The diversity of these species provides a means to examine the robustness of the methods against possible confounders such as rates of recombination, demographic effects, effective population size, and population structure. In cases where very large datasets (1000s of genomes) were available, we sub-sampled representative strains as described in Methods. A complete list of all isolates used in the analysis is given in Table S2.1.

For each species, we mapped the sequence reads to a single reference genome (Table 2.1), and defined alternative sets of core genes and IGRs using different frequency thresholds. Defining core gene sets on the basis that each core gene is universally present, or present at a very high frequency, among all the sequenced genomes is an established first step in bacterial

comparative genomics. This simplifies the analysis by removing the problem of missing data (genes), and by excluding mobile elements (e.g. phage and plasmids), and also reduces the problem of potentially conflicting signals resulting from high rates of recombination or atypical selection pressures. However, it is likely that the most frequently observed genes and IGRs that correspond to a strict core are also the most selectively constrained, thus this approach potentially imposes a bias. In order to address this, we analysed three different sets of core genes and IGRs for each species defined according to different frequency thresholds. The 'relaxed core' represents all genes and IGRs present in at least 2 genomes, the 'intermediate core' includes all genes and IGRs that are present in > 95% isolates, and the 'strict core', includes all genes and IGRs present in > 99% of isolates. The number of genes and IGRs included in each dataset is given in Table 2.1. The 'strict core' IGR dataset included at least 50% of the corresponding 'relaxed core' IGRs for each species, with the biggest potential bias in *S. pneumoniae* and *E. coli*, where the 'strict core' IGRs corresponded to 50.5% and 56.5% of the 'relaxed core' IGRs respectively. For the 'intermediate core' datasets, 64.9% and 64.2% of the relaxed core IGRs were included in the 'strict core' for *S. pneumoniae* and *E. coli* respectively.

Species	Data source	# Isolates	%GC	Reference genome	RC Genes	RC IGRs	IC Genes	IC IGRs	SC Genes	SC IGRs
E. coli	NCBI complete genome	157	50.8	MG1655	4305	3647	3164	2342	2873	2060
S. enterica	Public Health England	366	52.2	Typhimurium_D23580	4554	3777	3617	2830	3114	2456
K. pneumoniae	Holt et al, 2015	208	57.7	NTUH_K2044	4787	4006	3954	3150	2954	2453
S. aureus	Reuter et al, 2015	132	33.2	HO_5096_0412	2405	2084	2131	1704	2057	1609
S. pneumoniae	Chewapreecha et al, 2014	264	39.5	ATCC_700669	2183	1846	1574	1198	1373	932
M. tuberculosis	Casali et al, 2014	144	65.6	H37Rv	4069	3135	3806	2940	3332	2691

Table 2.1: The data used in the analysis. RC = Relaxed core, IC = Intermediate core, SC = Strict core.

Sequence properties of genes and IGRs

IGRs were identified based on reference genome annotation as described in Methods. The size distribution and GC content of both genes and assigned IGRs are shown in Figure 2.1. Figure 2.1a shows that IGRs with a predicted promoter at each end tend to be larger than double terminator regions and co-oriented regions. This is partly explained by the fact that many co-oriented regions are small spacers within operons. The GC content of IGRs is lower than in protein coding sequences (Figure 2.1b); although this difference is far less marked in *M. tuberculosis*, it is statistically significant in all species ($p < 10^{-16}$, Mann-Whitney *U* test).

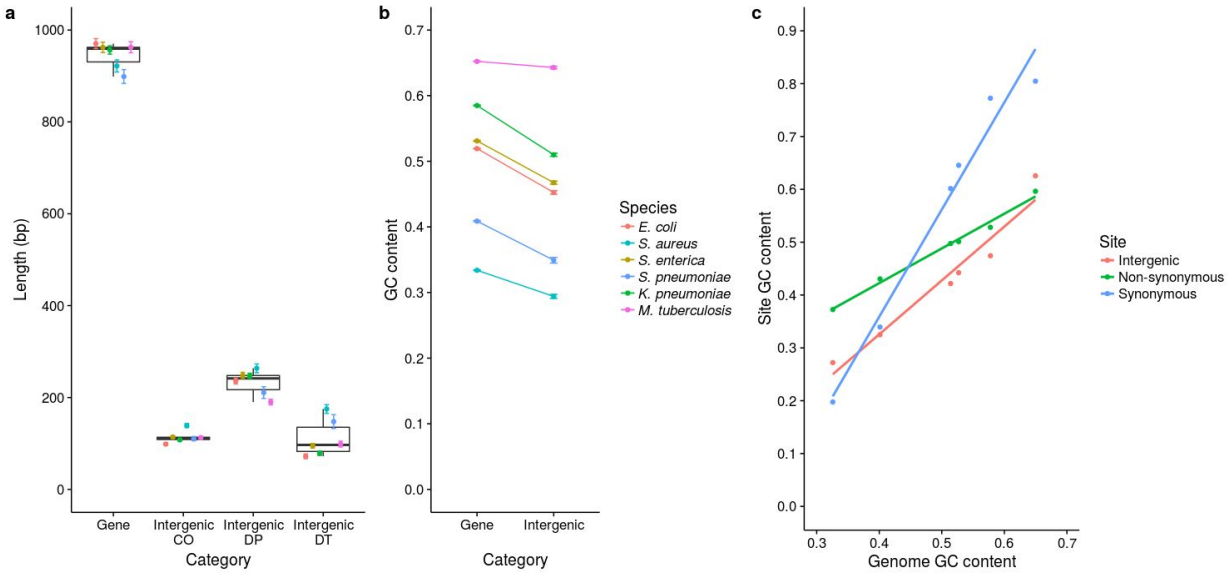

Figure 2.1: Summary of the sequence properties of genes and IGRs. a. Length distributions of genes and IGRs. IGRs were divided into three groups according to the orientation of the flanking genes: co-oriented regions are flanked by genes in the same orientation, double promoter regions are flanked by 5' gene starts, and double terminator regions are flanked by 3' gene ends. The points and error bars represent mean ± sem. **b.** GC contents of genes and IGRs. GC contents were calculated for each gene and IGR individually. The points and error bars represent the mean ± sem. **c.** GC contents of different site classes compared to genome GC content. The GC content of synonymous, non-synonymous, and intergenic sites was calculated, and compared with the genome GC content for each species. The steepness of the slope indicates the amount of constraint on the GC content of the site class (shallower slopes indicate stronger constraint).

Muto and Osawa showed that fourfold degenerate sites exhibit the widest range of GC content across a diverse sample of genomes (that is, these sites show the most extreme values), whereas non-degenerate second codon positions exhibit the narrowest range, with 1st and 3rd sites being intermediate. These authors noted that this variation in the range of GC content between different site categories mirrors the selective constraints on those sites, with second codon positions being the most constrained because they are in all cases non-degenerate

(Muto and Osawa 1987; Rocha and Feil 2010). We repeated this analysis using synonymous, non-synonymous and intergenic sites (Figure 2.1c). Our data are consistent with that of Muto and Osawa; the synonymous sites exhibit the widest range of GC content (steepest slope), the non-synonymous sites exhibit the narrowest range of GC content (shallowest slope). However, we also note that the slope for intergenic sites is intermediate between the synonymous and non-synonymous sites. If the original interpretation by Muto and Osawa is correct, this implies that the strength of selective constraint on intergenic sites is also intermediate between that on operating on synonymous and non-synonymous sites. Below we describe detailed analyses which examines this possibility in more detail.

The proportion of singleton mutations (PSM) is consistent with an intermediate strength of selective constraint on intergenic sites

In order to measure the frequency of strongly deleterious intergenic mutations, relative to synonymous, non-synonymous and nonsense mutations within coding regions, we used a simple method based on site frequency spectra. Similar methods have been used on non-coding DNA in eukaryotes (Drake et al. 2006) and in bacteria, albeit on a much smaller scale than the current study (Luo et al. 2011). Mutations affecting sites under strong selective constraint are more likely to be quickly purged by selection before they begin to rise in frequency, thus are also more likely to be very rare. Here, we define *very rare* mutations simply as those observed only once in the dataset (singletons). It is thus possible to gauge the proportion of strongly deleterious SNPs for a given site category simply by computing the proportion of those SNPs that are singletons (Proportion of Singleton Mutations; PSM). In order to check to what degree the definition of core IGRs imposes a bias we carried out the analysis using the three thresholds as defined above ('relaxed core', 'intermediate core' and 'strict core'). We first considered four mutation categories, intergenic, synonymous, non-synonymous, and nonsense. The PSM values for each of these mutation types, for all six species, are shown in Figure 2.2. An analysis of all individual genes and IGRs is given in Figure S2.2 ('intermediate core' only).

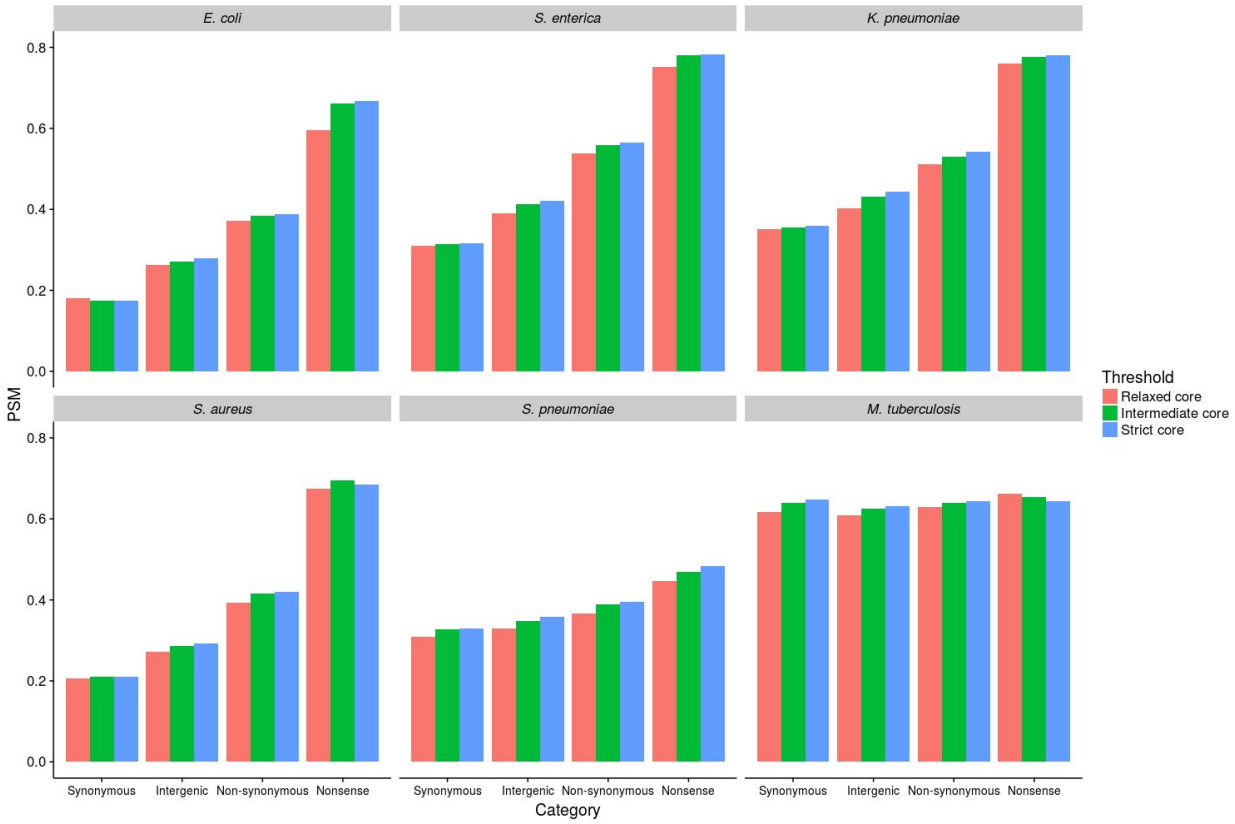

Figure 2.2: PSM (Proportion of Singleton Mutations) analysis of selection on different mutation categories. PSM values were calculated by dividing the number of singleton SNPs (those present in only one genome) by the total number of SNPs within that mutation category.

This analysis reveals a consistent trend across five of the six species, the exception being *M. tuberculosis*. Nonsense mutations had the highest PSM values, indicating the highest proportion of strongly deleterious mutations, followed by non-synonymous mutations, intergenic mutations, and finally synonymous mutations. Thus, for five of the six species, PSM values for intergenic sites were intermediate between the synonymous and non-synonymous PSM values. It follows that the proportion of SNPs at intergenic sites that are highly deleterious, and therefore purged rapidly by purifying selection, is intermediate between the equivalent proportions for synonymous and non-synonymous sites.

Although comparisons of PSM values between species are not valid, as species-specific factors (e.g. the rate of recombination) will also impact on PSM, it is reasonable to assume that these potential confounders are at least consistent between different mutation types within a single

species. This is strongly evidenced by the consistency of the relative strengths of selective constraint operating on each mutation type (nonsense > non-synonymous > intergenic > synonymous). The same trend is observed when individual genes and IGRs were analysed separately (Figure S2.2). Moreover, the pattern is highly robust to the definition of core genes are IGRs. Although the 'strict core' gene and IGR sets within each species correspond to marginally higher PSM values (as, expected and consistent with higher selective constraint), again the relative trends within each species remain robust. We are therefore confident that this analysis is not confounded by biases resulting from species-specific factors, or from selecting unrepresentative genes and IGRs. In *M. tuberculosis*, the exceptional species, there was very little difference between all mutation categories, and PSM scores were high in all cases. Multiple interpretations of the apparent patterns of selection and the high frequency of rare variants in *M. tuberculosis* have been discussed in the literature. These include very weak purifying selection, short coalescent time, linkage (background selection), a combination of purifying and positive selection, rapid demographic expansion combined with bottlenecks (leading to a reduction in the effective population size and increased drift), selective sweeps and diversifying selection (Namouchi et al. 2012; Pepperell et al. 2013; Hershberg et al. 2008). We consider some of these possibilities in the context of our results in more detail below.

We recognise that this analysis is potentially vulnerable to sequencing errors, as these are most likely to generate singleton SNPs. The consistency of the results across 5 diverse species is reassuring, as this is very difficult to reconcile with a high error rate without assuming this systematically affects some site categories more than others. Nevertheless, in order to gauge whether our analysis has been impacted by a high frequency of error-derived singleton SNPs, we repeated the analysis by first removing all singleton SNPs and instead computing, for each site category, the proportion of doubleton mutations (PDM). These are SNPs present in exactly two genomes within each sample; although still rare, these are *a priori* far less likely to have been generated by random sequencing error than singleton SNPs. PDM values were ordered nonsense > non-synonymous > intergenic > synonymous for all species except *M. tuberculosis* and *S. pneumoniae* (Figure S2.3). Thus, the only discrepancy between the PSM and PDM results was *S. pneumoniae*, where intergenic < synonymous. However, we note this discrepancy is marginal, and the distinction between site categories is less robust for this species even when considering PSM, probably reflecting very high rates of recombination in this species (discussed below).

To further examine to what extent singleton SNPs may have been generated by sequencing error, we carried out a detailed comparative analysis of the quality scores of singleton and non-singleton SNPs (Figure S2.4). We analysed three metrics: depth of coverage, proportion of reads supporting each variant, and the Phred Quality (Q) score in both singletons and non-singletons. The analysis revealed that the vast majority (> 99%) of all SNPs were of extremely high quality, and that there are negligible differences in the quality scores between singleton and non-singleton SNPs. For example, across all species 99.5% of singleton SNPs and 99.8% of non-singleton SNPs had a Q score of > 100. This quality score corresponds to an error rate of 10^{-10}, or equivalently one erroneous SNP every 2000 genomes (given a 5Mb genome). Given these combined checks, we are highly confident that errors in the singleton SNPs have not confounded our analysis.

The signal of purifying selection on intergenic sites is time-dependent

To further examine selective constraint on intergenic sites, we extended the logic of dN/dS by computing dI/dS, where dI = intergenic SNPs per intergenic site. dI has previously been used in *M. tuberculosis* as a neutral reference by calculating dI/dS for individual IGRs using neighbouring genes as a source of synonymous sites (Wang and Chen 2013). In contrast, we drew pairwise comparisons by pooling sites across the whole genome, and used the genome-wide dI as the numerator and the genome-wide dS as the denominator. We computed genome-wide dN/dS in the same way in order to draw valid comparisons between the strength of selection on intergenic sites and non-synonymous sites, both relative to synonymous sites.

Previous work has shown that dN/dS decreases with divergence time due to a lag in purifying selection, which operates much more strongly on non-synonymous than synonymous sites as the former are more likely to be slightly deleterious (Rocha et al. 2006; Namouchi et al. 2012). We tested for the same time dependence in dI/dS by comparing pairs of very closely related genomes (within 'clonal complexes' (CCs); where dS < 0.003) with those representing more distantly related genomes ('between-CCs'; dS > 0.003. This analysis was also carried out for all three alternative gene sets (relaxed, intermediate and strict core; Figure 2.3). All genome comparisons within *M. tuberculosis* were defined as 'within-CC' due to the very low level of variation in this species. We also plotted, for each pair of isolates and for each species, dN/dS and dI/dS against dS in order to further explore the impact of divergence time on dI/dS

('intermediate core' only; Figure S2.5).

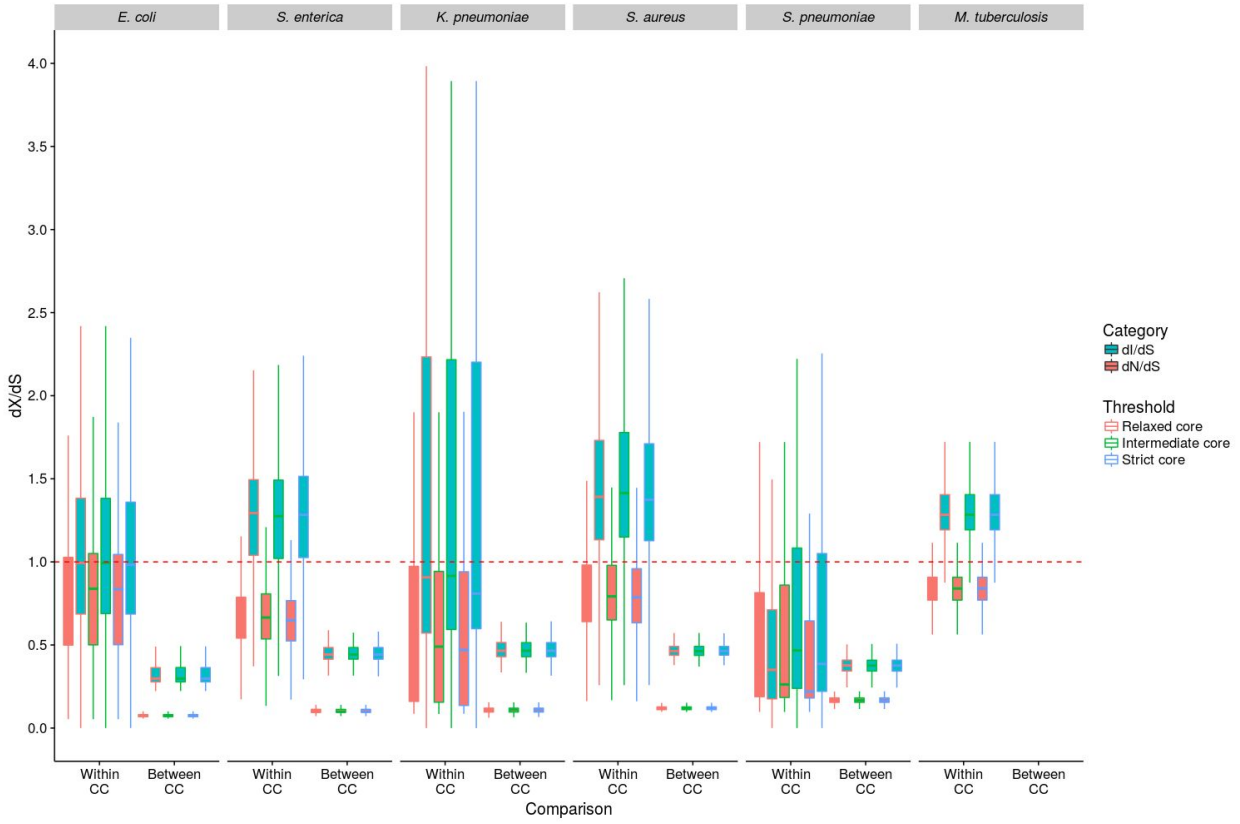

Figure 2.3: dN/dS and dI/dS analysis of selection. dN/dS and dI/dS were calculated between isolates in a pairwise manner. The results were categorised into within clonal complex (Within CC, dS < 0.003, red), and between clonal complex (Between CC, dS > 0.003, blue) comparisons, to account for the effect of divergence time on the observed levels of selection. The notches in the box plots represent 95% confidence intervals around the median. All comparisons between *M. tuberculosis* isolates were classified as 'Within CC' due to the extremely low level of diversity in this species. The dashed red line shows where dN/dS and dI/dS = 1, and therefore indicates neutrality.

Figure 2.3 shows that for each species, dI/dS is consistently greater than dN/dS for both within and between-CC comparisons. The between-CC dN/dS and dI/dS values are universally < 1, but the within-CC values are more equivocal, with the dN/dS values being mostly < 1, and the dI/dS values being < 1 in *E. coli, K. pneumoniae,* and *S. pneumoniae*, and > 1 in *S. enterica, S. aureus,* and *M. tuberculosis*. This trend is consistent across all three core gene and IGR sets, with very little difference in dI/dS and dN/dS values between the sets. Low dN/dS and dI/dS values (particularly in the between-CC comparisons) are strong evidence of purifying selection on non-synonymous and intergenic sites, and lower dN/dS values compared to dI/dS values indicate stronger constraint on non-synonymous sites than intergenic sites. It is worth noting

that this observation (dI < dS) has previously been interpreted as evidence for positive selection on synonymous sites (Wang and Chen 2013). Given previous work and the results of the PSM analysis, we argue instead that it confirms greater selective constraint on intergenic sites than on synonymous sites. Moreover, the difference between within and between-CC comparisons is evidence of time dependence consistent with ongoing purifying selection operating over increasing divergence, as noted previously for non-synonymous sites (Rocha et al. 2006). For four of the five species for which such a comparison was possible, (*E. coli*, *S. aureus*, *S. enterica*, *K. pneumoniae*), dN/dS and dI/dS were both significantly higher for within-CC comparisons than between-CC comparisons ($p < 10^{-16}$, Mann-Whitney *U* test). These differences are expected if non-synonymous and intergenic SNPs are preferentially purged (relative to synonymous SNPs) over divergence time, although we recognise that our pairwise methodology might lead to an amplification of these differences due to the over-sampling of long internal branches in the between-CC comparisons. In *S. pneumoniae*, dN/dS was significantly higher for within-CC comparisons compared to between-CC comparisons ($p < 10^{-16}$, Mann-Whitney *U* test) but dI/dS was not ($p = 0.19$). It is possible that the signal of time dependence is weaker in this species owing to high rates of recombination (Chaguza et al. 2016). The statistical analysis described above was carried out based on the 'intermediate core' gene set. As there was negligible difference between the three core gene and IGR sets in both the PSM and dI/dS analyses presented thus far, we performed all subsequent analyses on the 'intermediate core' sets (where at least 90% of genes and IGRs are present in at least 95% of isolates).

We also plotted, for each pair of isolates for each species, dN/dS and dI/dS against dS (based on the 'intermediate core' only; Figure S2.5). In the case of *E. coli*, *S. aureus*, *S. enterica* and *K. pneumoniae* a large number of points are evident at very low values of dS; these reflect the presence of clusters of closely related genomes in these species (i.e. clonal complexes). The absence of significant clonal clustering in *S. pneumoniae* reflects high rates of recombination, and can help to explain the lack of significant difference within and between clonal complexes in this species as noted above. However, for all species except *S. enterica* and *M. tuberculosis* there is a significant decrease of both dN/dS and dI/dS against dS ($p < 10^{-16}$, Spearman's correlation). The time dependence of dN/dS potentially poses a problem for comparing between species, as those species with longest time to most recent common ancestor will appear to be under stronger selection (as dN/dS decreases with divergence time). However, Figure S2.5

shows that dN/dS decreases very quickly initially, and then begins to plateau very early, suggesting that this is not a major problem in our analysis.

As noted above, the genetic diversity within *M. tuberculosis* is so low that between-CC comparisons were not possible, as the dS for all pairwise comparisons was < 0.003. Values of dN/dS are approximately 0.8 in this species, which is comparable to the within-CC values for all the other species, and slightly higher than that reported previously for this species (Pepperell et al. 2013; Hershberg et al. 2008). Both the observation of high dN/dS in *M. tuberculosis* and the PSM analysis described above are consistent with, though not demonstrative of, weak purifying selection in this species. It has been argued that weak purifying selection in *M. tuberculosis* reflects its lifestyle as an obligate pathogen subject to frequent bottlenecks, and thus a reduction in effective population size (Hershberg et al. 2008). Contrary to this view, Namouchi *et al.* highlighted the absence of the classic footprints of genome degradation expected to result from increased drift, and that, in contrast to Hershberg *et al.* 2008, non-synonymous SNPs are more common (compared to synonymous SNPs) in terminal branches of the tree (as predicted if purifying selection is operating). Moreover, other authors have reported evidence of both positive and purifying selection (Pepperell et al. 2013; Farhat et al. 2013), and even diversifying selection in *M. tuberculosis* (Osório et al. 2013). Given this complex picture, it is possible that our results represent a mixture of contrasting forces acting over short coalescence times, converging on a signal that is indistinguishable from very weak purifying selection. In favour of this argument, the diversity in *M. tuberculosis* is so low that only a small number of mutations would need to be positively selected in order to have a large impact on the patterns observed, and our *M. tuberculosis* sample is enriched for antibiotic resistance which is known to be positively selected (Farhat et al. 2013; Casali et al. 2014; Pepperell et al. 2013).

Purifying selection on intergenic sites is strongest near gene borders

Although values of dI/dS < 1 are consistent with stronger selective constraint on intergenic sites than on synonymous sites, this could also arise due to slower mutation rates within IGRs than in coding regions. This might be expected if a non-negligible fraction of mutations arose during transcription, which would also impact on intergenic sites near to the gene border (Chen and Zhang 2013). The demonstration of the time dependence of dI/dS, specifically the difference between within and between-CC comparisons, acts to mitigate these concerns, but as a further check we calculated dI/dS values from intergenic sites immediately upstream of genes (30

bases upstream from the start codon). If intergenic sites immediately upstream of genes are transcribed, and transcription-derived mutation significantly elevates dS, then dI/dS should approach 1 in these regions. However, for each species (except *M. tuberculosis*) we noted the opposite; dI/dS immediately upstream of genes was in fact lower than dI/dS for intergenic sites in general ($p < 10^{-16}$, Mann-Whitney *U* test), suggesting that transcription-derived mutation is not confounding our analysis (Figure S2.6). This suggests that intergenic sites close to the start of genes are under particularly strong purifying selection, which may be due to the presence of regulatory elements upstream of genes, or selection for mRNA stability to enable efficient translation (Molina and Van Nimwegen 2008).

The strength of purifying selection on different classes of intergenic regulatory element

Above we demonstrate that intergenic sites in the majority of bacterial species are likely to be under selective constraint. However, we have not yet considered to what extent the strength of purifying selection may vary within a given IGR according to the presence or absence of different regulatory elements. It would be expected that sites associated with known or predicted regulatory elements should be under stronger selective constraint than sites with no known function, and it may be the case that certain classes of regulatory element are under stronger selective constraint than others. To test this possibility, we identified all ribosome binding sites (RBSs), non-coding RNAs, predicted promoters, and rho-independent terminators for each species (see Methods). We then applied both methods (PSM and dI/dS) to compare the strength of selective constraint on these different elements, as well as on all the remaining intergenic sites that do not correspond to any of these elements ('unannotated sites').

With the exception of *M. tuberculosis*, we note that the PSM values for the RBSs tend to be higher than for other regulatory elements and unannotated sites (Figure S2.7), suggesting that these elements are particularly strongly constrained. In *E. coli*, *S. aureus*, and *K. pneumoniae*, non-coding RNAs also appear to be strongly constrained. In contrast, promoters and terminators tend to exhibit similar PSM values to the unannotated sites. We next drew the same comparisons using dI/dS values (Figure 2.4). This confirmed the observation from the PSM analysis of particularly strong purifying selection on RBSs in all species except *M. tuberculosis*, and in non-coding RNAs of *E. coli*, *S. aureus*, and *K. pneumoniae*. Indeed, the dI/dS values for RBSs and non-coding RNAs in these species are similar to the dN/dS values, suggesting that the strength of purifying selection on these elements is similar to that operating on

non-synonymous sites (Figure 2.4). This analysis also reveals that predicted promoters and terminators tend to be under more similar levels of selective constraint to unannotated sites. The two analyses (PSM and dI/dS) are highly concordant, with both showing a clear signal of strong purifying selection on RBSs and non-coding RNAs in the same species, and that promoters and terminators are under similar levels of purifying selection to unannotated sites. Importantly, however, it is clear that (with the exception of *M. tuberculosis*) dI/dS is < 1 in all cases, including unannotated sites, which suggests a high level of constraint even when excluding major regulatory elements.

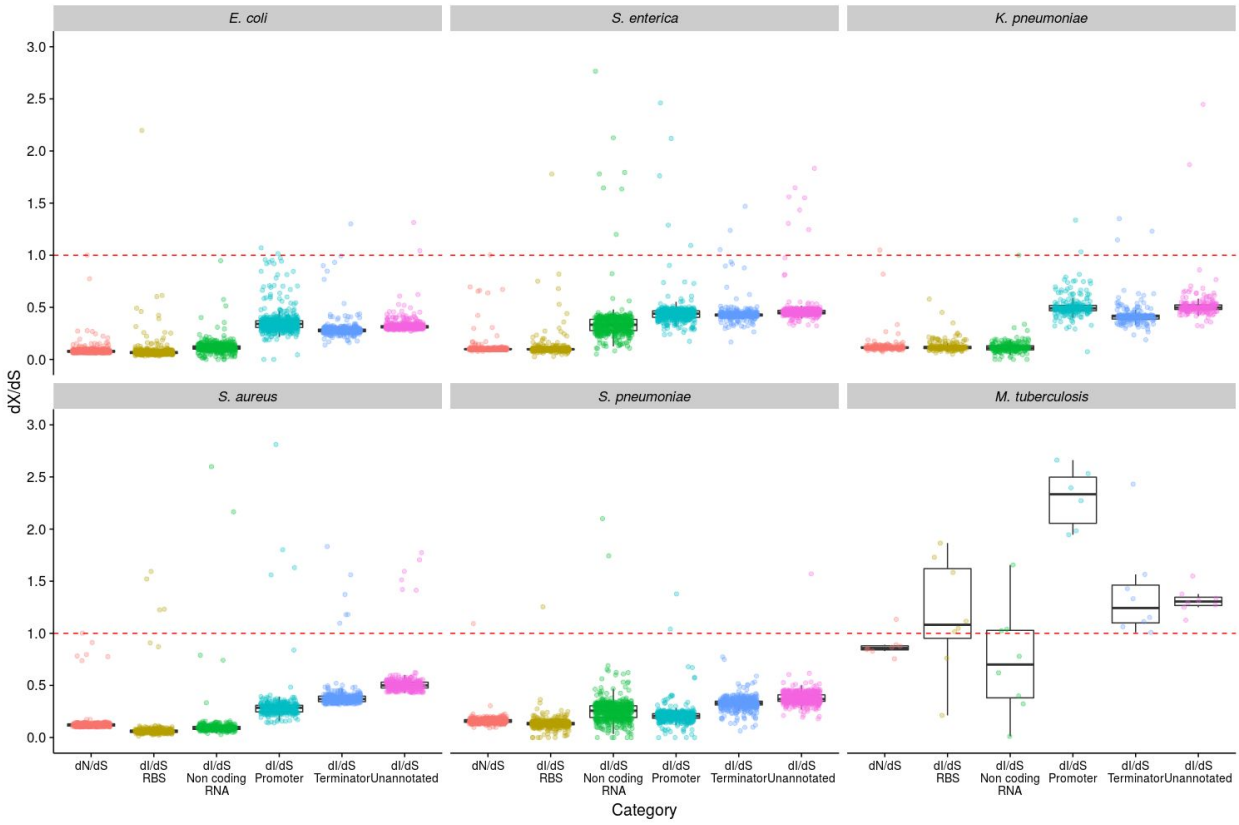

Figure 2.4: dI/dS analysis of selection on different regulatory elements. dI/dS was calculated between isolates in a pairwise manner, and the results were binned by dS (bin width = 0.0001) to control for oversampling of very closely related isolates (such as those belonging to the same CC). The genome-wide dN/dS values are included to enable comparisons to be made between non-synonymous sites and the different regulatory intergenic sites. The dashed red line shows where dN/dS and dI/dS = 1, and therefore indicates neutrality.

We then further examined the strength of selective constraint on transcriptional terminators which appear to be under only marginally stronger selective constraint than unannotated sites.

Transcriptional terminators consist of a stem-loop structure, and it seemed likely that the stem should be under stronger constraint than the loop, due to requirement of the stem sequence to maintain complementary base pairing. To test this, we calculated dI/dS for the terminator stems and loops separately (Figure S2.8). As expected, the stem dI/dS values are substantially lower than those for the loop, confirming that the stem is more constrained than the loop, and providing additional validation of our methodology. However, we also note that the dI/dS values for the loops are clearly < 1 in *S. aureus*, *E. coli* and *S. pneumoniae*, indicating they are not completely free to change in these species.

As discussed, our analysis points to considerable selective constraint (relative to synonymous sites) on intergenic sites even when the major regulatory elements are excluded. We noted earlier (Figure S2.6) that the strength of selective constraint appears to be particularly high within 30-bp of the gene borders. In order to examine to what extent this trend reflects the presence of known regulatory elements, we first excluded these elements then investigated SNP density as a function of the distance from gene start codons in co-oriented IGRs (where the genes flanking these regions are in the same orientation). In each species (except *M. tuberculosis*), SNP densities increased with distance from the 5' gene starts ($p < 10^{-4}$, Spearman's correlation) (Figure 2.5), demonstrating that the relatively high level of selective constraint on intergenic sites near gene borders (noted earlier) remains even when excluding promoters, terminators, RBSs and non-coding RNAs.

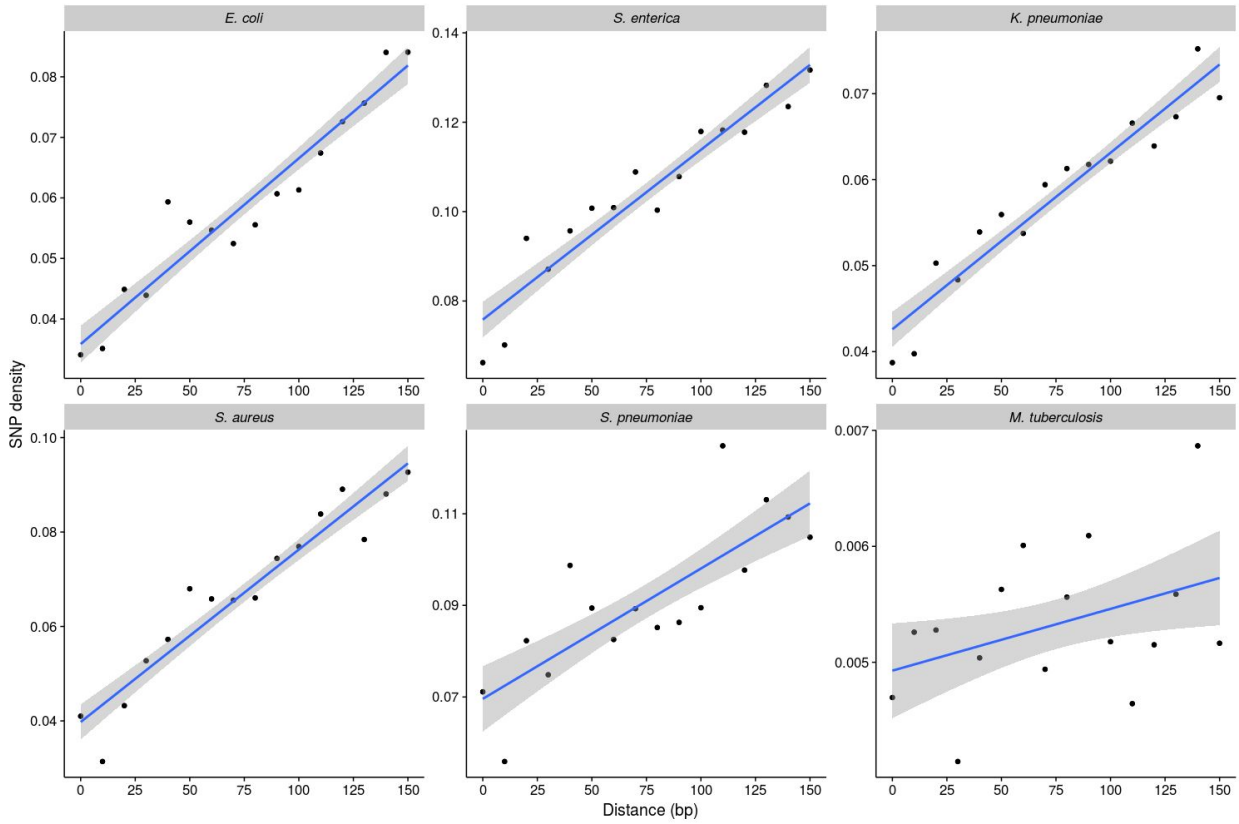

Figure 2.5: Analysis of SNP densities within co-oriented IGRs (those flanked by genes in the same orientation as each other). SNP densities were calculated in 10 bp windows moving away from the gene start codon by dividing the number of SNPs by the number of IGRs of that length or greater (to normalise for the unequal lengths of IGRs). Only unannotated intergenic sites were considered in the analysis.

Evidence for positive selection within IGRs of *M. tuberculosis*

Throughout this analysis, *M. tuberculosis* has repeatedly proved the exception as it exhibits very little evidence of purifying selection on protein-coding sequences, and even some evidence of positive selection on IGRs. Considering different intergenic regulatory elements separately reveals that positive selection is strongly associated with predicted promoter regions. The mean dI/dS for *M. tuberculosis* promoters was 2.8 (Figure 2.4), and the vast majority (97%) of comparisons exhibit a dI/dS of > 1 (Figure S2.9). This result is not solely a consequence of the approach we have used to calculate dI/dS, which corrects for mutation biases and base composition, as even without this correction the mean dI/dS for promoters is 1.9. The evidence for positive selection is highly statistically significant. Of the 10513 promoter sites, 99 have experienced a SNP, compared with 6330 of the 954745 synonymous sites ($p < 0.0001$ by a Fisher's exact test). We further confirmed significance by resampling the predicted promoter and

synonymous sites 1000 times and comparing the distributions with a z-test ($p < 10^{-16}$). Figure 2.3 shows that the within-CC values of dI/dS for *S. enterica* and *S. aureus* are > 1, thus are also indicative of positive selection. We therefore also calculated dI/dS separately for the different intergenic elements for all the other 5 species, but restricting the analysis to within-CC comparisons. This did not reveal any evidence of positive selection on promoters or any other IGR elements (Figure S2.10).

In order to further investigate the potential functional relevance of promoter SNPs in the *M. tuberculosis* dataset, we identified genes downstream of predicted promoters harbouring SNPs (Table S2.2). The 71 promoter SNPs identified corresponded to 58 genes; 11 genes were identified for which the corresponding promoter harboured 2 SNPs, and one gene where the promoter harboured 3 SNPs. Many of the downstream genes are known to play a key role in virulence, resistance or global regulation. For example, 8 genes were transcriptional regulators, and promoters in four of these experienced two independent SNPs: MT0026 (a putative HTH type regulator); *CmtR* (a cadmium sensing repressor (Chauhan et al. 2009)), *WhiB2* and *WhiB4* (transcription factors (Larsson et al. 2012; Ma et al. 2015; L. J. Smith et al. 2012)). Six genes were members of the PE/PPE protein family that are recognised virulence factors (Fishbein et al. 2015). Genes known to play a role in resistance are also identified, including one mutation in the *ethA* promoter; mutations in this promoter have previously been implicated in resistance to ethionamide (Casali et al. 2014). The promoter for the alanine dehydrogenase gene *ald* is the only example harbouring 3 independent SNPs, loss of function of this gene has recently been shown to confer resistance to D-cycloserine (Desjardins et al. 2016). Other notable genes include *ctpJ*, which encodes an ATPase that controls cytoplasmic metal levels (Raimunda et al. 2014), and *psk2* which plays a critical role in the synthesis of cell wall lipids (Sirakova et al. 2001). In addition, 15 hypothetical genes residing downstream of mutated promoters were identified, and in five of these cases the promoter experienced two independent SNPs.

Discussion

Here we demonstrate consistent evidence for purifying selection on intergenic sites in five diverse species (excluding *M. tuberculosis*), even when major regulatory elements are excluded. This further challenges the view that IGRs can be used as mostly neutral markers to estimate neutral mutation rates or profiles (Wang and Chen 2013). Rather, our results suggest these regions are rich with functional elements, many of which are yet to be characterised, and

are selectively conserved and maintained (Molina and Van Nimwegen 2008; Degnan, Ochman, and Moran 2011; Luo et al. 2011). Although consistent with previous work, this observation is pertinent with respect to the default exclusion of IGRs from bacterial databases based on the core genome (cg)MLST model (M. C. J. Maiden et al. 2013; Jolley and Maiden 2010; Sheppard, Jolley, and Maiden 2012; M. C. J. Maiden and Harrison 2016). Our analysis suggests the exclusion of IGRs from these databases is not warranted either for biological nor technical reasons.

We have used two fast and efficient approaches to measuring selection on non-protein coding sequences based on established principles of population genetics and suited for large whole genome datasets. According to the nearly-neutral theory, slightly deleterious mutations are not eliminated immediately from a population, but can persist for a period of time determined by the selection coefficient (s) and the effective population size (N_e) (Ohta 1973; Kimura and Ohta 1971). Highly deleterious mutations will be lost more quickly whilst they are still very rare. The rarest SNPs are those that are observed in only one genome (singletons), thus the proportion of singleton mutations (PSM) reflects the frequency of highly deleterious mutations (Hershberg et al. 2008). The weaker effect mutations will tend to be lost more gradually over time (Rocha et al. 2006). Whereas the PSM approach provides a measure of how many SNPs are purged very quickly due to highly deleterious effects, dI/dS provides a measure of how many deleterious mutations have been purged relative to the coalescence time of the genomes under consideration. Thus, these two methods are not only independent but also provide complimentary comparisons encompassing both strongly and more weakly deleterious mutations.

We demonstrate for the first time that, like dN/dS (Rocha et al. 2006; Castillo-Ramírez et al. 2011), dI/dS also decreases with divergence time, as the ratio is lower when considering between (rather than within) CC comparisons. This confirms that the lower prevalence of segregating sites in IGRs when compared to synonymous sites (i.e. dI/dS < 1) does not simply reflect differences in mutation rate, and moreover our analysis of IGR sequences near gene borders reveals that dS has not been significantly inflated by transcription-derived mutation. We also note that our analyses are likely to be conservative. The comparator (dS) is not a perfect neutral benchmark; selection at synonymous sites operates on codon usage bias (Sharp et al. 2005), secondary RNA structure (Molina and Van Nimwegen 2008), and possibly GC content

(Hildebrand, Meyer, and Eyre-Walker 2010; Balbi, Rocha, and Feil 2009; Rocha and Feil 2010; Namouchi et al. 2012). Moreover, dI/dS and dN/dS will continue to decrease with divergence time until the synonymous sites are saturated, and there is no reason to suppose that the available data corresponds to the minima for a given species.

Our analyses provides a novel comparison of the strength and direction of selection on different classes of regulatory element within IGRs. This reveals that RBSs and non-coding RNAs tend to be under relatively strong constraint, broadly comparable to non-synonymous sites. We have shown that the average selection operating on terminator regions reflects strong selection on the stem, combined with much weaker selective constraint on the loop. Our results also demonstrate that purifying selection is operating on IGRs (relative to synonymous sites) even when predicted promoters, terminators, RBSs and non-coding RNAs are excluded, and that this constraint is strongest close to gene starts. This suggests that many functional elements in IGRs remain uncharacterised, and unannotated intergenic sites close to gene borders may have particular functional significance.

The power of our approach is underscored by novel evidence for positive selection in predicted promoter regions in *M. tuberculosis*. This result is highly statistically significant, meaning that the signal of positive selection must be strong enough not to be confounded by any background purifying selection in our global comparisons. In order to gauge the functional relevance of these promoter SNPs, we identified all downstream genes, and noted a number global regulators, transcription factors, and genes implicated in virulence or resistance. A large number of hypothetical proteins were also identified, which could form targets for future studies (Table S2.2). This observation thus points to a key role of subtle changes within promoters for short-term adaptation through regulatory rewiring in this species, which may also help to account for the paucity of variation within coding regions. A recent report by McNally *et al.* is consistent with this view as it implicated a key role for changes in promoters within a single *E. coli* clone (ST131) as an adaptive response coinciding with the gain and loss of accessory elements (McNally et al. 2016).

Conclusion

Here we have applied two tests to quantify the strength and direction of selection acting on IGRs in bacteria. We demonstrate consistent evidence of strong purifying selection on IGRs in 5

species, even when major regulatory elements are excluded. We also note the strength and direction of selection varies with the class of intergenic regulatory element, the species under consideration and distance from gene border. We show that the signal of purifying selection increases with divergence time for intergenic sites, just as it does for non-synonymous sites and consistent with expectations under the nearly-neutral model of evolution. Although our analysis is consistent with previous reports of very weak purifying selection in *M. tuberculosis* (Hershberg et al. 2008), we are cognisant that this evidence is equivocal and that our data may in fact reflect a complex combination of purifying, positive and possibly diversifying selection operating over short coalescence times (Pepperell et al. 2013; Farhat et al. 2013; Casali et al. 2014; Osório et al. 2013). In support of this, we note strong evidence for positive selection within *M. tuberculosis* promoters, and argue that regulatory rewiring represents a major adaptive mechanism in this species.

We conclude that our current understanding of the functions encoded in IGRs is fragmented, and we would therefore urge utmost caution before excluding these regions from bacterial databases or 'core' genome analyses. Our results call for the routine analysis of the selection pressure operating on, and hence functional relevance of, IGRs similar to those carried out on protein-coding regions. To facilitate this, the code used in the analysis is available at https://github.com/harry-thorpe/Intergenic_selection_paper under the GPLv3 license.

Chapter 3

Piggy: A Rapid, Large-Scale Pan-Genome Analysis Tool for Intergenic Regions in Bacteria

The work presented in this chapter is available as a preprint at:

Thorpe, Harry A., Sion C. Bayliss, Samuel K. Sheppard, and Edward J. Feil. 2017. "Piggy: A Rapid, Large-Scale Pan-Genome Analysis Tool for Intergenic Regions in Bacteria." bioRxiv. doi:10.1101/179515.

Commentary text

The work in this chapter builds on the previous chapter by incorporating intergenic sites into pan-genome analyses. This is facilitated by the development of a tool, Piggy, which enables analyses of both gene and intergenic components of bacterial pan-genomes. *E. coli* and *S. aureus* are used as example datasets, and for *S. aureus* these are combined with RNA-seq data to show that changes in intergenic regions affect gene expression. The statement of authorship for this chapter can be found in the Appendix, supplementary form SF2.

Abstract

Despite overwhelming evidence that variation in intergenic regions (IGRs) in bacteria impacts on phenotypes, most current approaches for analysing pan-genomes focus exclusively on protein-coding sequences. To address this we present Piggy, a novel pipeline that emulates Roary except that it is based only on IGRs. We demonstrate the use of Piggy for pan-genome analyses of *Staphylococcus aureus* and *Escherichia coli* using large genome datasets. For *S. aureus*, we show that highly divergent ('switched') IGRs are associated with differences in gene expression, and we establish a multi-locus reference database of IGR alleles (igMLST; implemented in BIGSdb). Piggy is available at https://github.com/harry-thorpe/piggy.

Introduction

Whole-genome sequencing has revealed that, in many bacteria, individual strains frequently recruit new genes from a seemingly endless genetic reservoir. The total complement of genes observed across all strains, known as the pan-genome, often numbers tens of thousands, up to an order of magnitude more than the number of genes present in any single genome. In contrast, the 'core-genome', which refers to the complement of genes present in all (or the vast majority) of sampled isolates, can be significantly smaller than the total number of genes in any given genome (Medini et al. 2005; Page et al. 2015). For example, a study of 328 *Klebsiella pneumoniae* isolates, each of which harbour 4-5,000 genes, revealed a pan-genome of 29,886 genes; only 1,888 (6.8%) of which were universally present (core) (Holt et al. 2015). Similarly, genome data for 228 *Escherichia coli* ST131 isolates revealed a pan-genome of 11,401 genes, of which 2,722 (23.9%) were core (McNally et al. 2016). The degree of gene content variation in the latter study is particularly striking as these isolates were all from the same sequence type (ST), thus show limited nucleotide divergence in core genes, and are descended from a recent common ancestor.

There is growing recognition that the acquisition of new genes through horizontal gene transfer (HGT) has a central role in ecological adaptation (Vos et al. 2015). The emergence and spread of antibiotic resistance, underpinned by the transfer of plasmids and other MGEs, is a pertinent example. The increasing availability of datasets containing thousands of isolates thus offers an unprecedented opportunity for describing the genetic basis of bacterial adaptation. However, the scale of these data presents serious logistic and conceptual challenges in terms of data management and analysis.

Pioneering pan-genome analysis tools, such as PanOCT and PGAP relied on all-vs-all BLAST comparisons between protein sequences, and scaled approximately quadratically with the number of isolates (Fouts et al. 2012; Zhao et al. 2012). LS-BSR introduced a pre-clustering step which substantially reduced the number of BLAST comparisons, but sacrificed specificity (Sahl et al. 2014). More recently, the Roary pipeline has rapidly gained in popularity for scalable, user-friendly, pan-genome characterisation (Page et al. 2015). Roary uses a pre-clustering step based on CD-HIT (L. Fu et al. 2012), and is more accurate and faster than LS-BSR, meaning that it can analyse 1000s of isolates quickly using modest computing resources.

The concept of the pan-genome, as described above, places an exclusive emphasis on genes; or, more specifically, open reading frames with the potential to encode proteins. This gene-centric perspective has both shaped, and been shaped by, the bioinformatics tools developed to interrogate the pan-genome. For example, Roary works by taking individual protein-coding sequences, pre-defined using Prokka annotation (Seemann 2014), and assigning each to a single cluster of homologous sequences. This approach thus excludes non protein-coding intergenic regions (IGRs) which typically account for approximately 15% of the genome. This is clearly problematic for downstream attempts to identify genotype-phenotype links, as IGRs contain many important regulatory elements including, but not limited to, promoters, terminators, non-coding RNAs, and regulatory binding sites. Moreover, we have recently shown that IGRs are subject to purifying selection in the core-genomes of diverse bacterial species, even when known major regulatory elements are excluded (Thorpe et al. 2017; Molina and Van Nimwegen 2008).

Given that variation in IGRs can have profound phenotypic consequences, it is timely to consider how best to incorporate these sequences into pan-genome analyses. A key question is the degree to which protein-coding genes, and their cognate regulatory elements, should be considered a single 'unit', both selectively (in terms of co-adaptation) and in terms of physical linkage on the chromosome. If physical linkage is assumed to be highly robust, such that genes are mostly transferred along with their cognate IGRs, then in principle the definition of a 'gene' could be expanded to include the upstream regulatory regions. On the other hand, if there is moderate or weak linkage between genes and IGRs, such that IGRs can occasionally transfer

independently, then the purview of the pan-genome could be expanded to include the full complement of IGR alleles in addition protein-coding sequences.

Consistent with the second model, which allows for independent transfer of IGRs, a landmark study demonstrated that *E. coli* genes can apparently be regulated by alternative IGRs that frequently share no sequence similarity to each other (Oren et al. 2014). Moreover, the distribution of these IGRs was incongruent with gene trees, suggesting that recombination can act to replace one IGR with another resulting in regulatory 'switches'; a process they call horizontal regulatory transfer (HRT) (Oren et al. 2014). It is important to note here that the term ‘switching’ refers only to the replacement of an IGR by a non-homologous or highly divergent variant sequence. It does not specify that the replacement IGR has a particular origin, and could therefore correspond to a transfer from elsewhere in the same genome, or from another isolate. It was also noted that conserved flanking genes may facilitate this process by providing localised regions of homology. IGR switches can be accompanied by differential gene expression (Oren et al. 2014), and may provide a mechanism to offset the fitness costs of harbouring plasmids and other MGEs (McNally et al. 2016), pointing to a central role for this process in adaptation.

Our current understanding of the evolutionary dynamics of IGRs in the context of bacterial pan-genome leave many open questions. Specifically, it is unclear how IGRs are distributed among isolates within bacterial populations, how commonly IGRs and their cognate genes are co-transferred, or how the frequency of HRT relates to different functional gene categories. A more complete understanding of bacterial adaptation clearly requires a careful consideration of gene presence/absence alongside gene regulation. Here we address this by introducing a new pipeline called Piggy which closely emulates and complements the established pan-genome analysis pipeline Roary (Page et al. 2015). Input and output files for Piggy and Roary use the same format, and run in a similar time on modest computing resources. Piggy provides a means to rapidly identify IGR switches, and more broadly the means to examine the role of horizontal transfer in shaping the bacterial regulome. We demonstrate the utility of Piggy using large genome datasets for single lineages within two bacterial species, both of which are of high public health importance; *Staphylococcus aureus* ST22 (EMRSA-15) and *Escherichia coli* ST131. Conventional pan-genome analyses are applied to analyse and compare core and accessory IGRs/genes in these lineages. In *S. aureus* we show a link between IGR switching

and changes in gene expression, and demonstrate proof-of-principle by establishing a multilocus IGR scheme, (igMLST) in BIGSdb (Jolley and Maiden 2010). Piggy is available at (https://github.com/harry-thorpe/piggy) under the GPLv3 licence.

Methods

Datasets

The *S. aureus* ST22 dataset was assembled from published genome sequences of the clinically important lineage ST22 (EMRSA-15) (Reuter et al. 2015) available at http://www.ebi.ac.uk/ena (study number ERP001012). The original genome assemblies were used, and 500 isolates belonging to ST22 were randomly selected for analysis. The *S. aureus* RNA-seq data was previously published (Warne et al. 2016), and is available at (http://www.ebi.ac.uk/ena, study number ERP009279). This was supplemented with the corresponding reference genomes, HO_5096_0412: HE681097, MRSA252: BX571856, Newman: AP009351, S0385: AM990992, available at (www.ncbi.nlm.nih.gov). The *E. coli* ST131 dataset was also from a previously published study (McNally et al. 2016), and is available at (http://datadryad.org/resource/doi:10.5061/dryad.d7d71). All complete genomes and assemblies were annotated with Prokka (Seemann 2014).

Roary and Piggy parameter settings

Roary (Page et al. 2015) was run using default parameters except for the following: -e -n (to produce alignments with MAFFT (Katoh and Standley 2013)); -i 90 (lower amino acid identity than the default); -s (to keep paralogs together); -z (to keep intermediate files). Piggy was run using default parameters except for --len_id, which controls the percentage of IGR sequences which must share similarity in order to be clustered together. For the *S. aureus* ST22 and *E. coli* ST131 datasets, Piggy was run twice, once with --len_id 10 and once with --len_id 90. The former was used for the pan-genome comparisons between genes and IGRs (Figs 2 and 3) in order to be comparable with Roary. Using a low length identity (--len_id 10) enabled homologous sequences of varying lengths (for example a truncated sequence) to cluster together. Roary does not provide a similar setting, and only requires that sequences have a minimum length of 120 bp. It is common that genes in Roary clusters frequently vary considerably in length (likely due to both genuine differences and assembly artefacts), and are clustered together despite this. Thus, in order to provide a fair comparison between Roary and Piggy (and not increase the number of IGR clusters due to strict clustering), a relaxed --len_id

setting of 10 was used. The latter (--len_id 90) was used whenever 'switched' IGRs were detected, as this enabled more control over downstream filtering of these sequences.

RNA-seq analysis

Two biological replicates for each isolate were analysed. Kallisto (Bray et al. 2016) was used to quantify transcripts (--kmer-size 31 and --bootstrap-samples 100), and Sleuth (Pimentel et al. 2017) was used to normalise and filter the counts produced by Kallisto. These counts were then $\log_{10}$ transformed, and major axis (MA) regression was performed. Rockhopper2 (Tjaden 2015) was used to produce an operon map for each strain by grouping adjacent genes with similar expression profiles together into operons.

Statistical analysis

All statistical analysis was performed within R version 3.3.2 (https://www.r-project.org). All plotting was performed with ggplot2 (Wickham 2009).

Results

Overview of the Piggy pipeline

Figure 3.1a shows an overview of the Piggy pipeline. The first step is to run Roary, as the gene presence absence output file from Roary is used as an input for Piggy. Piggy is then run using the same annotated assemblies as Roary, specifically GFF3 format files such as those produced by Prokka (Seemann 2014). Piggy extracts intergenic sequences (IGRs) from these files, and uses the flanking gene names and their orientations to name the IGRs (Figure 3.1b). Each IGR name contains three pieces of information: the upstream gene, the downstream gene, and their relative orientations (CO - co-oriented, DP - double promoter, DT - double terminator). For example, the IGR ‘Gene_1 Gene_2 DP’ is flanked by Gene_1 and Gene_2, which are divergently transcribed away from each other. For IGRs at the edge of contigs the missing information is denoted by NA, for example ‘Gene_1 NA NA’. Including the gene neighbourhood information gives context to the IGR and enables identification of 'switched' IGRs. The IGRs are then clustered with CD-HIT (L. Fu et al. 2012) at user defined identity thresholds (--nuc_id - nucleotide identity, --len_id - length identity). The nucleotide identity is defined as SNPs/aligned sites, and the length identity is defined as shared sites/alignment length. These two flags allow the user to set the level of stringency for clustering. For example, a conservative approach is to set high values for both nucleotide and length identity such that IGRs must be similar in both

nucleotide and length identity to cluster together. By relaxing the length identity whilst maintaining a high nucleotide identity threshold, highly related sequences still cluster even if one is truncated. A representative sequence from each cluster is then used to perform an all-vs-all BLASTN search (Camacho et al. 2009). This is used to merge similar clusters, which did not cluster with CD-HIT. These clusters are then used to produce an IGR presence absence matrix ('IGR_presence_absence.csv'), in the same format as the gene presence absence matrix ('gene_presence_absence.csv') produced by Roary. Up until this point, the pipeline is very similar to Roary (Page et al. 2015).

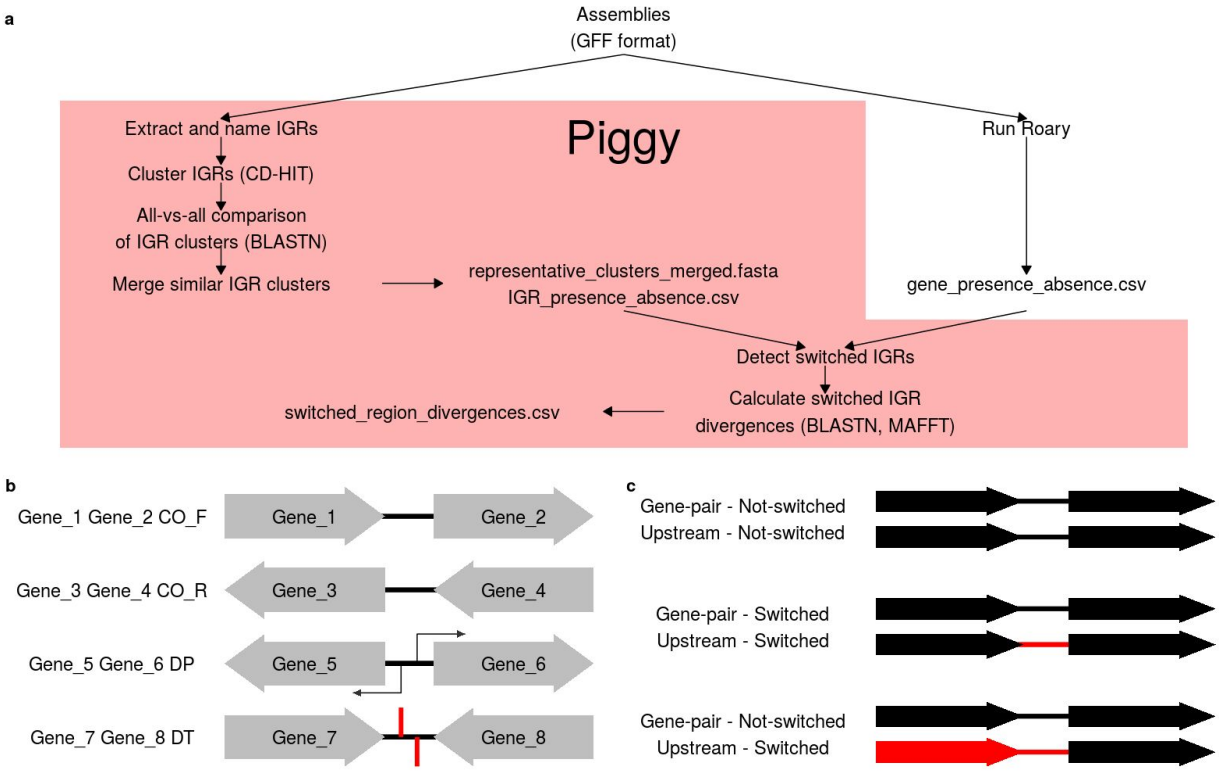

Figure 3.1: An overview of the Piggy pipeline. **a.** A schematic to illustrate the Piggy pipeline and how it works alongside Roary. **b.** IGRs are named according to their flanking genes and their orientations. This naming scheme enables Piggy to link genes with their associated IGRs, and provides information on their orientations. **c.** A schematic to illustrate the difference between the 'gene-pair' and 'upstream' methods used to identify candidate switched IGRs.

Switched IGR detection

Piggy identifies 'switched' IGRs using two methods. For both methods, the term ‘switch’ refers to divergent IGR sequences adjacent to genes. This definition does not specify a particular origin for the divergent IGR sequences, in keeping with (Oren et al. 2014). The first method identifies adjacent genes on the same contig (gene-pairs), and searches for IGR clusters which lie between these gene-pairs (Figure 3.1c). Instances where multiple IGR clusters correspond to

the same gene-pair are identified as candidate switched IGRs. The second method identifies instances where multiple IGR clusters are upstream of the same gene, which are also putatively switched IGRs. This is a less conservative approach as the downstream gene is not considered in this case, (Figure 3.1c). The gene-pair method is used by default as it controls against detecting 'switching' (recombination) events that encompass more than a single IGR, for example, cases where a mobile element has inserted between two genes. However such cases remain relevant as the regulation of the downstream gene will still be affected.

To ensure that differences in gene annotation between isolates are not erroneously identified as 'switching' events, the first and last 30 bp of each flanking gene are searched against the IGRs with BLASTN. Any matches from these searches indicate differences in annotation of gene borders (rather than genuine differences between the IGRs), and these sequences are disregarded. In order to confirm that they represent genuine switching events, candidate switched IGRs are searched against each other with BLASTN with low complexity filtering turned off (-dust no). If there is no significant match they are classed as 'switched', and if there is a significant match they are aligned using MAFFT (Katoh and Standley 2013). The resulting alignment is then used to calculate nucleotide identity (SNPs / aligned sites), and length identity (number of shared sites / alignment length). These values can then be used to define an appropriate threshold to identify 'switched' IGRs. To aid this, Piggy calculates within-cluster divergences for both genes and IGRs, and these divergences can be used to calibrate Piggy with Roary.

***Staphylococcus aureus* ST22**

In order to validate Piggy, we ran it on a dataset of 500 *S. aureus* ST22 isolates. *S. aureus* ST22 (EMRSA-15) is a clinically important hospital-acquired methicillin resistant strain which is common in the UK and is rapidly expanding elsewhere in Europe and globally. Previous work has shown that *S. aureus* ST22 is clonal and has a relatively small set of accessory genes (Holden et al. 2013; Reuter et al. 2015). The size of the gene and IGR pan and core-genomes were compared by running 500 ST22 (Reuter et al. 2015) isolate genomes through Roary and Piggy. Frequency histograms and accumulation curves were plotted for both genes and IGRs (Figure 3.2).

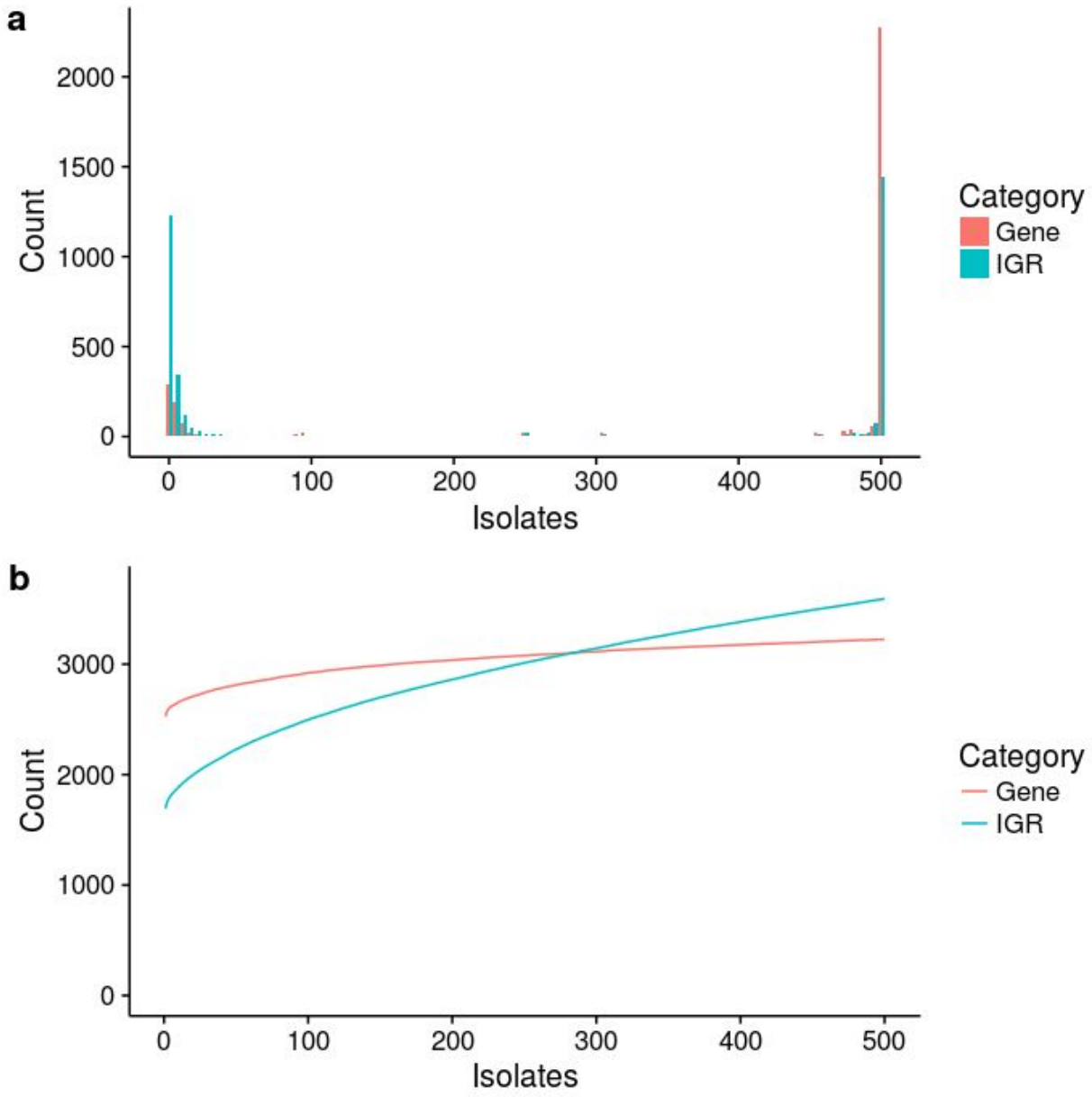

Figure 3.2: Properties of the *S. aureus* ST22 pan-genome. Genes (red) and IGRs (blue) were analysed. **a.** Gene and IGR frequency histogram – that is, the number of genes / IGRs present in any given number of isolates. The vast majority of genes / IGRs are either very rare or very common. **b.** Gene and IGR accumulation curves – that is, the cumulative number of genes / IGRs detected in a given number of isolates.

The gene-IGR frequency histogram (Figure 3.2a) shows that there are 2,312 core genes and 1,486 core IGRs, where core is defined as gene presence in > 99% of isolates. The fact that there are fewer core IGRs than core genes is in part due to the exclusion of intra-operonic IGRs < 30 bp. Both distributions conform to the U-shape typically found in such analyses, where the majority of genes/IGRs are either very common or very rare. The gene accumulation curve

(Figure 3.2b) shows a total of 3,225 genes, with a mean of 2,524 genes per isolate. The gradient of the curve is shallow, consistent with the small, closed, pan-genome of clonal ST22 isolates. The IGR curve shows that each isolate has fewer IGRs than genes (1,696 on average per isolate) due to the exclusion of IGRs < 30 bp, but that the total number of IGRs (3,593) is higher than the total number of genes reflecting greater diversity in IGRs than genes. The IGR curve increases more steeply than the gene curve, and does not appear to plateau. Despite these differences, within any given isolate on average 92% of genes and 88% of IGRs were core.

***Escherichia coli* ST131**

The utility of Piggy was further validated by re-analysing data from a recent study on the widespread and clinically important *E. coli* lineage ST131 (McNally et al. 2016). This dataset contains 236 clinical *E. coli* ST131 isolates from human, domesticated animal, and avian hosts. *E. coli* is a more genetically diverse species than *S. aureus*, and unsurprisingly *E. coli* ST131 has a larger pan-genome than *S. aureus* ST22, with 12,806 genes and 16,429 IGRs (Figure 3.3a). Of these, 3,285 genes and 1,403 IGRs were core (Figure 3.3b), out of an average of 4,678 genes and 2,999 IGRs per isolate.

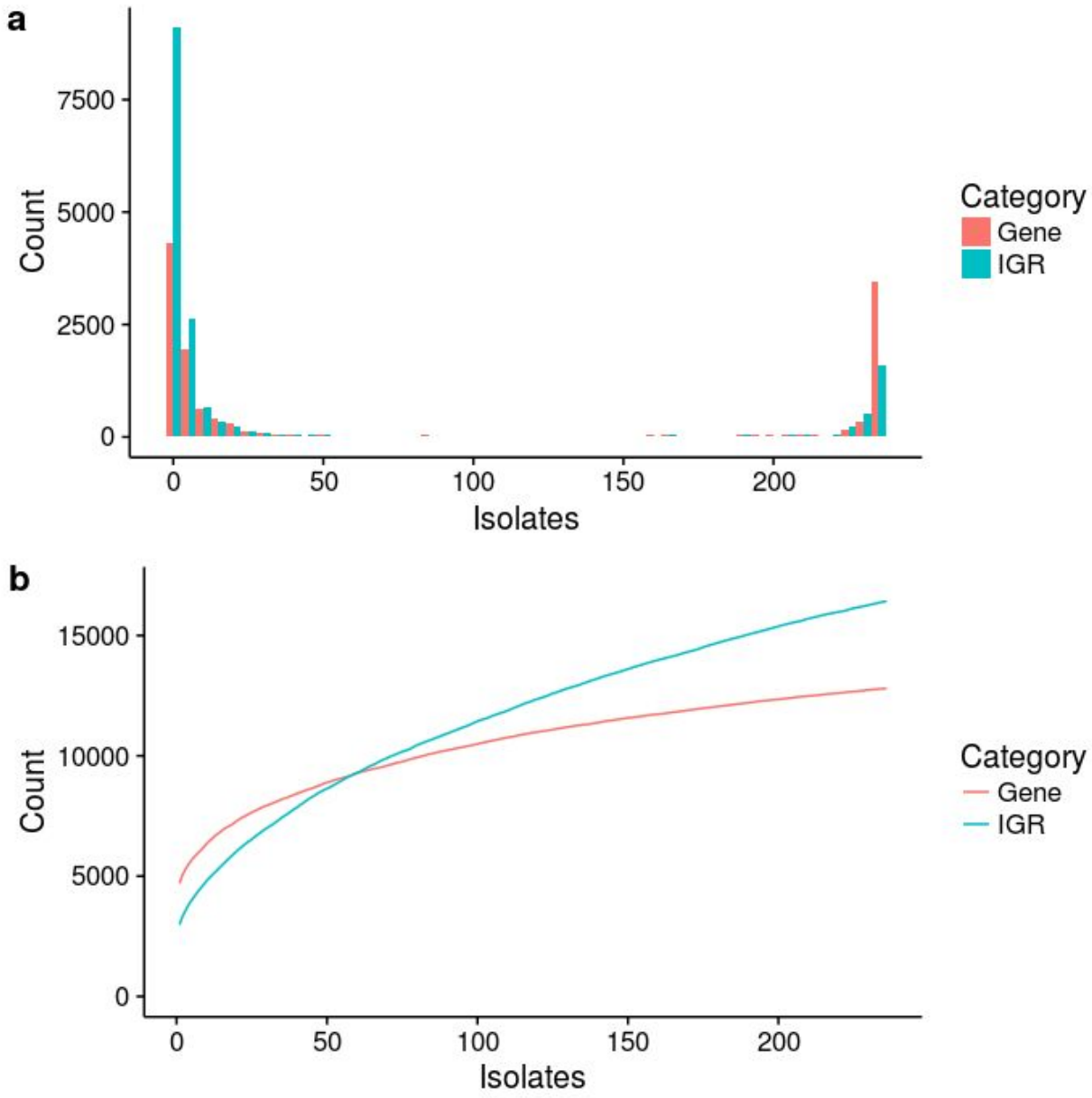

Figure 3.3: Properties of the *E. coli* ST131 pan-genome. Genes (red) and IGRs (blue) were analysed. **a.** Gene and IGR frequency histogram – that is, the number of genes / IGRs present in any given number of isolates. The vast majority of genes / IGRs are either very rare or very common. **b.** Gene and IGR accumulation curves – that is, the cumulative number of genes / IGRs detected in a given number of isolates.

Thus despite the differences in diversity, for both *S. aureus* and *E. coli* datasets we found a lower number of core IGRs than core genes, but a high number of accessory IGRs compared to accessory genes. This is illustrated by the fact that the IGR and gene accumulation curves intersect in both species. A lower proportion of both genes (70%) and IGRs (47%) are core within each *E. coli* ST131 isolate, compared to *S. aureus* ST22. Similarly, rare accessory genes

and IGRs are much more prevalent in *E. coli* ST131 than in *S. aureus* ST22 with 34% of genes and 55% of IGRs found in < 1% of isolates in *E. coli* ST131, compared with 11% of genes and 40% of IGRs in *S. aureus* ST22.

Previous work has found evidence of extensive IGR switching, where the linkage between an IGR and the cognate downstream gene breaks down, resulting in alternative gene / IGR pairs (Oren et al. 2014). Piggy provides a list of candidate switching events together for both 'gene-pair' and 'upstream' approaches (see Methods) at different thresholds of nucleotide identity. For the *E. coli* ST131 data, the pipeline detected 61 cases of putative IGR switching using the most conservative settings (i.e. the conservative gene-pair method, and the alternative IGRs showing no sequence similarity by BLASTN). Relaxing the threshold of sequence identity to < 90% resulted in the identification of an additional 317 candidate switching events, though these possibly reflect either relaxed or positive selection.

Switched IGRs influence gene expression in *S. aureus*

To examine whether switches in IGRs affect the expression of cognate (downstream) genes, we used a previously published RNA-seq dataset based on four reference *S. aureus* isolates HO_5096_0412 (ST22), Newman (CC8), MRSA252 (CC36), and S0385 (CC398) (Warne et al. 2016). Each of these *S. aureus* references isolate represents a distinct major clonal complex, and all were grown under identical conditions with each experiment being replicated. Thus these data provide evidence of the natural variation in gene expression within the *S. aureus* population. By analysing these data alongside the output from Piggy, it is possible to test the extent to which IGR switches between these four genomes can account for the observed variation in gene expression between clonal complexes. First Roary was used to identify a set of 2094 single copy core genes present in all four isolates, and then expression of these core genes was quantified using Kallisto (Bray et al. 2016). To do this we used RNA-seq data for two replicates for each of the four reference genomes. We then used Sleuth (Pimentel et al. 2017) to normalise and filter these counts.

To check the consistency of the data between biological replicates, we first plotted two replicates for each isolate against each other (e.g. Newman replicate 1 vs Newman replicate 2) (Figure 3.4). These plots were tightly correlated (mean R^2 = 0.98), confirming that the expression values for individual genes were consistent between replicates. We then plotted

between-isolate comparisons, again using both replicates for each genome (e.g. Newman replicate 1 vs MRSA252 replicate 1, and Newman replicate 2 vs MRSA252 replicate 2) (Figure 3.4). These comparisons revealed considerably more scatter, with R^2 values ranging from 0.76 to 0.9. Given the extremely high R^2 values for within-isolate comparisons, the decrease in R^2 for between-isolate comparisons reflects genuine differences in expression between the isolates. We note that a small number of genes show very striking differences in expression between the clonal complexes. For example, the expression of *mepA*, which encodes a multidrug efflux pump, was ~250 fold higher in Newman compared with the other isolates.

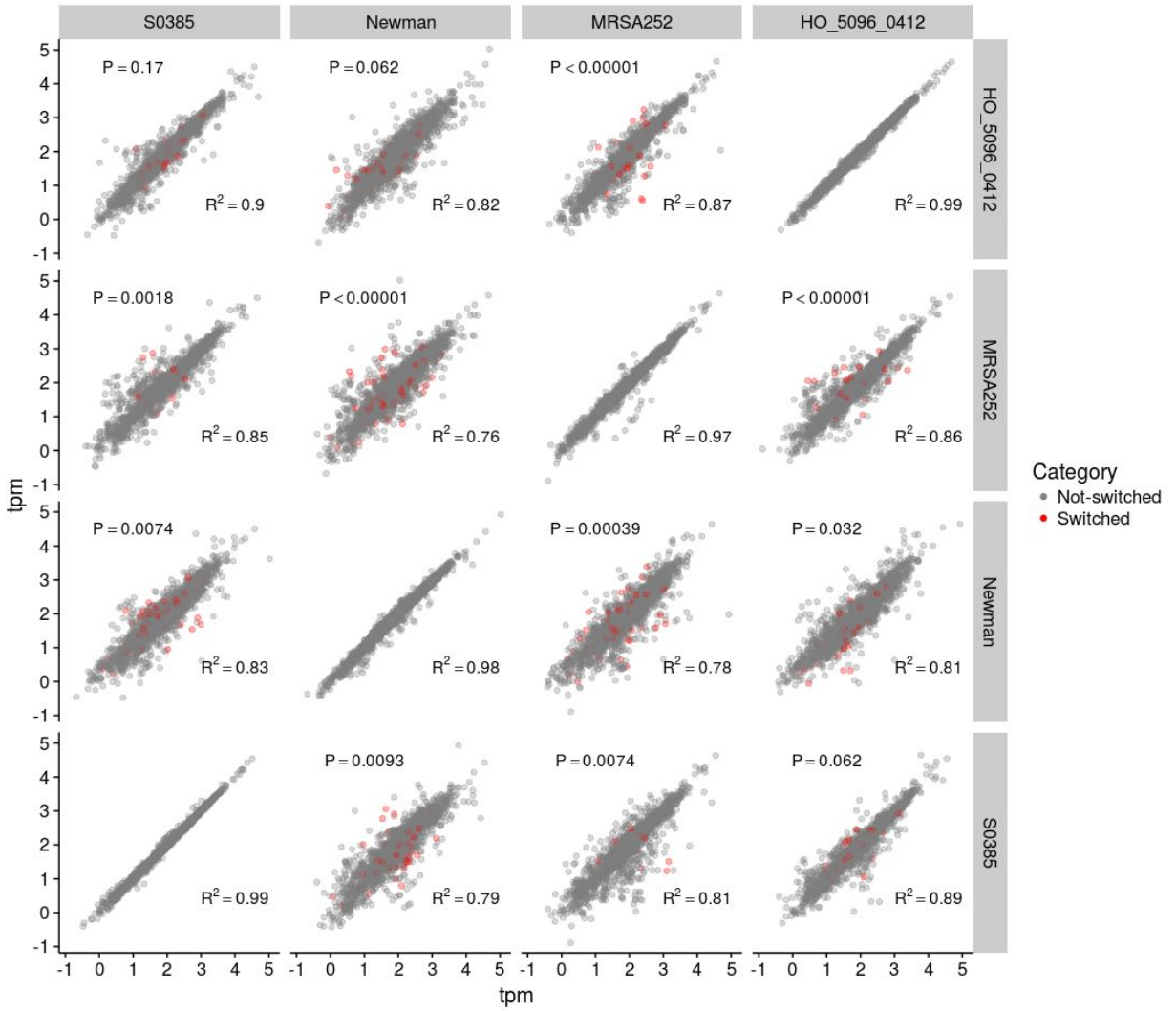

Figure 3.4: *S. aureus* gene expression data. Pairwise RNA-seq comparisons between four *S. aureus* isolates, where two biological replicates were used for each isolate. The top-left of the diagonal corresponds to comparisons between replicate 1 from different isolates (e.g. SO385 replicate 1 vs HO_5096_0412 replicate 1). The bottom-right of the diagonal corresponds to comparisons between replicate 2 from different isolates (e.g. SO385 replicate 2 vs HO_5096_0412 replicate 2). The diagonal corresponds to comparisons between the two biological replicates from the same isolate. 2094 core genes were analysed in each comparison, and tpm (Transcripts per Kilobase Million) was used to quantify expression. The genes were separated into two categories: Switched (red), and Not-switched (grey), based on their upstream IGRs. The R^2 value corresponds to all the genes. The P-value corresponds to a Monte Carlo permutation test comparing the residuals of the two groups of genes, where a significant score indicates that the genes downstream of switch IGRs are associated with a greater degree of differential expression (i.e. greater residuals).

The genomes of each pair of isolates were analysed using Roary and Piggy to identify switched IGRs with a nucleotide identity threshold of < 90% for IGR clusters. For each pair of isolates, we

then identified all genes immediately downstream of a switched IGR. As a single switched IGR might impact on the expression of more than one co-transcribed downstream genes we also considered all genes linked in a single operon that could be impacted by a single switching event upstream affecting a shared promoter. Thus, for each pair of isolates we identified all core genes putatively affected by upstream IGR switches. We then tested whether these genes showed a higher degree of differential expression by conducting Monte Carlo permutation tests on the residuals from the regressions (Figure 3.4). For each pairwise comparison of isolates, we summed the residuals of the genes with switched IGRs (shown as red points in Figure 3.4), and compared this to a distribution obtained by resampling (without replacement) 100,000 random sets of the same number of genes and summing their residuals. We computed a one-tailed p-value by dividing the number of permutations with summed residuals greater than the observed value by 100,000. We then adjusted the p-values using the Benjamini-Hochberg method (Figure 3.4). Because we used both replicates separately (e.g. Newman replicate 1 vs S0385 replicate 1, and Newman replicate 2 vs S0385 replicate 2), each comparison between pairs of isolates was tested twice independently. In 9/12 pairwise comparisons, the observed residuals of the genes downstream of switched IGRs were significantly greater than expected from the resampled data, indicating that genes with switched IGRs were more differentially expressed than those without. Of the three remaining comparisons, two corresponded to comparisons between HO_5096_0412 and S0385 ($P = 0.17$, and $P = 0.062$), and one between HO_5096_0412 and Newman ($p = 0.062$). The second comparison between HO_5096_0412 and Newman was the most weakly significant result ($p = 0.032$). Thus, the two replicates for each individual pairwise comparison were largely concordant with each other.

Our analysis confirms that genes downstream of switched IGRs are on average more likely to be differentially expressed than genes not associated with IGR switches as identified using Piggy. To illustrate the genomic context and expression differences of genes with switched IGRs, we selected three of the most differentially expressed genes with IGR switches for the Newman vs MRSA252 comparison, and plotted nucleotide identity across the IGR (calculated as a 20-bp sliding window) alongside gene expression (Figure 3.5).

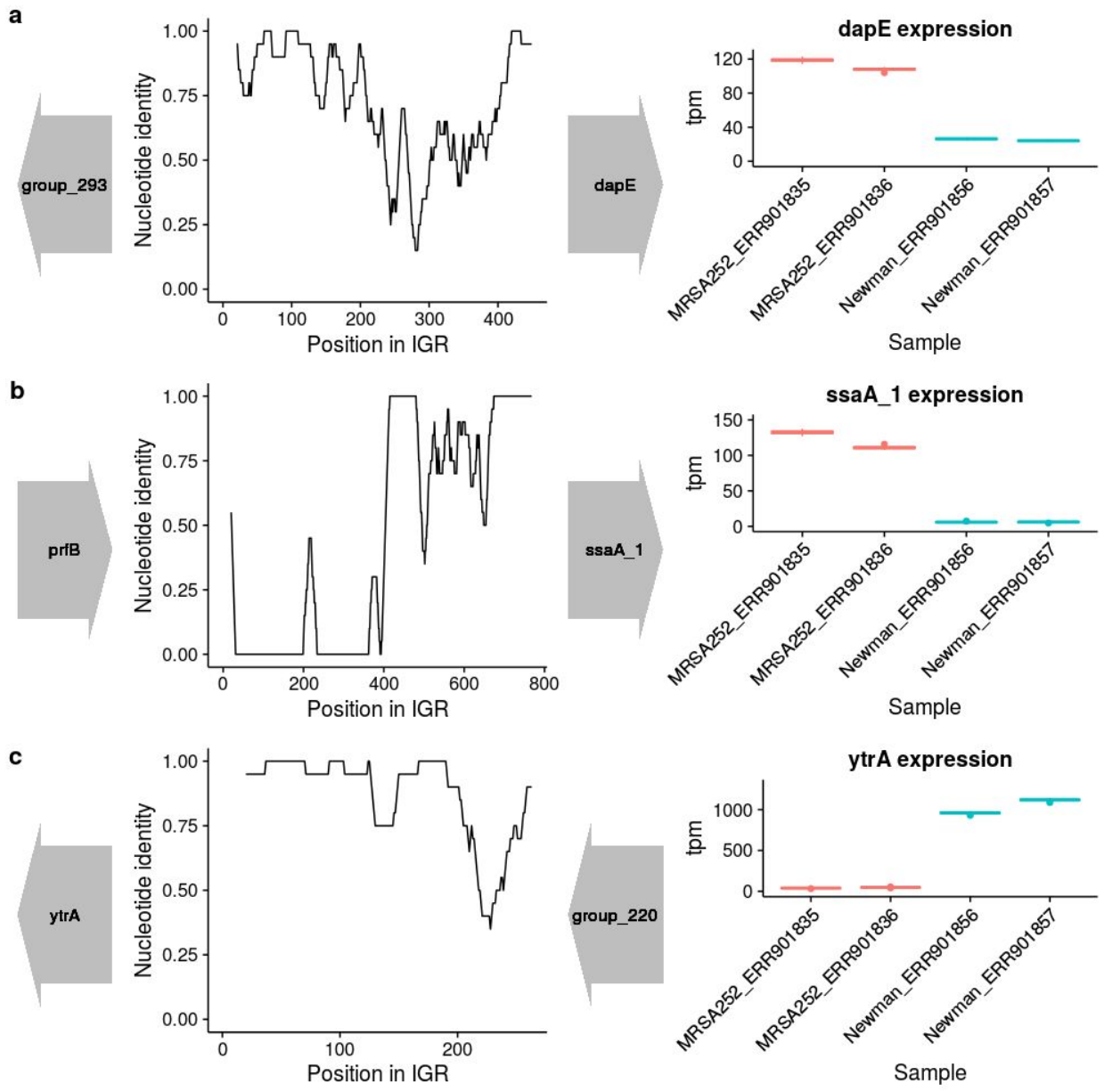

Figure 3.5: A detailed view of the genomic neighbourhood and expression data for selected genes in Newman vs MRSA252. Nucleotide identity was calculated using a 20 bp sliding window across the IGR, and this is shown alongside the flanking genes in their correct orientation (left). The corresponding expression data for the gene of interest was also shown (right), with the two boxplots per isolate corresponding to the two biological replicates. **a.** dapE **b.** ssaA_1 **c.** ytrA.

Compatibility and scalability

We have so far demonstrated that Piggy can be used to analyse the intergenic component of the pan-genome and identify IGR switches, and shown that these switches have biological relevance with respect to gene expression. Importantly, Piggy is designed such that the output

files are compatible with existing software and databases. The 'IGR_presence_absence.csv' file has an identical format to the 'gene_presence_absence.csv' file produced by Roary, and can be loaded directly into the interactive browser-based viewer Phandango (Hadfield et al. 2017) (Figure S3.1). It can also be used as input, along with a traits file, to Scoary (Brynildsrud et al. 2016) to test for associations between IGRs and phenotypic traits. Moreover, the 'representative_clusters_merged.fasta' file can be loaded directly into BIGSdb (Jolley and Maiden 2010) to create an allele scheme for IGRs. In order to provide proof-of-principle, we created a multilocus IGR (igMLST) scheme in BIGSdb. Briefly, 2631 unique IGR sequences with length ≥ 30bp, from 7 *S. aureus* reference genomes, were entered into the database locus list. Using functionality within the database, these sequences were grouped as a searchable scheme (S_aureus_Intergenic_PIGGY), comparable to MLST, rMLST and wgMLST schemes (M. C. J. Maiden et al. 2013; Jolley et al. 2012; Sheppard, Jolley, and Maiden 2012). The distribution of IGRs was analysed for all isolates in the database, identifying IGRs as present in the respective genome if a hit was recorded with nucleotide identity ≥ 70% over ≥ 50% of the sequence using a BLAST word size of 7 bp. The scheme can be found at https://sheppardlab.com/resources. [N.B. The BIGSdb IGR scheme was generated by Sion Bayliss, University of Bath, Bath, UK.] Finally, Piggy runs in a comparable time to Roary and scales approximately linearly with increasing numbers of isolates, as tested on a MRC-CLIMB (Connor et al. 2016) virtual machine with 10 vcpus and increasing numbers of *S. aureus* ST22 isolates (Figure S3.2).

Discussion

Whole-genome sequence datasets consisting of hundreds or even thousands of bacterial isolates have revealed pan-genomes of many thousands of genes and large differences in gene content between isolates of the same species. Currently, pan-genome diversity is considered almost exclusively in terms of protein-coding genes, despite overwhelming evidence that variation within IGRs impacts on phenotypes. Here we address this by introducing Piggy, a pipeline specifically designed to incorporate IGRs into routine pan-genome analyses by working in close conjunction with Roary (Page et al. 2015).

The utility of this approach is demonstrated using large datasets of *S. aureus* ST22 and *E. coli* ST131. Consistent with previous analyses of protein-coding regions (Holden et al. 2013; McNally et al. 2016), the IGR component of the ST131 pan-genome (the 'panIGRome') is

considerably larger than that for ST22. There was more diversity within IGRs than genes in both species. While some IGRs may be essential for expression of multiple genes, it is expected that IGRs will be subject to less stabilizing selection than protein coding genes (Thorpe et al. 2017). The maintenance of core IGRs in both bacterial genome datasets is consistent with selection acting to conserve them and allows alignment and analysis in much the same way as protein-coding regions.

Variation within regulatory elements located within IGRs can impact on the expression of the downstream gene (Oren et al. 2014). Piggy (alongside Roary) provides the means to combine information on genes and their cognate IGRs thus facilitating the detection of 'switched' IGRs and downstream genes that are potentially affected. We have shown that in *S. aureus*, genes with switched upstream IGRs show a higher degree of differential expression than those without. This is consistent with previous work on *E. coli* (Oren et al. 2014), and suggests that the identification of IGR switches using Piggy can provide a useful indication of differential gene expression, even in the absence of RNA-seq data. However, we note that high divergence within IGRs does not necessarily imply selection for differential gene expression, and may instead simply reflect weaker selective constraints. A clear direction for future work is to make constructs consisting of genes with alternative IGRs, in order to directly measure the effect of natural IGR variants on gene expression. Similar experiments have previously been performed in *E. coli* based on variation within promoters (Shimada et al. 2014), and IGRs more broadly (Oren et al. 2014).

Excluding IGRs from bacterial comparative genomics severely limits our ability to draw inferences on the regulation of gene expression and associated phenotypic consequences. By developing Piggy as an easy-to-use bioinformatics tool with output files that are compatible with existing software and databases (eg Roary, Phandango; Figure S3.1, Scoary, BIGSdb) we envisage that combined information from genes and their cognate IGRs will vastly improve our understanding of genome evolution in bacteria.

Chapter 4

Variation in deleterious mutation load in *H. pylori* populations, and effect of selection on introgressed DNA in hpEurope

Introduction

Helicobacter pylori is a gram-negative Proteobacterium which lives in the human stomach. It is thought to be carried by half of humans worldwide, and it causes gastric inflammation in all carriers, gastric ulcers in 10-15% of carriers, and gastric carcinoma in ~1% of carriers (Kodaman et al. 2014; Maixner et al. 2016). Transmission of *H. pylori* mostly occurs between family members, and because carriage is often long-term, this results in strong phylogeographic structure within the population (Moodley et al. 2012). *H. pylori* has been associated with humans for approximately 100,000 years, and phylogeographic patterns of *H. pylori* resemble major migration events in human history (Moodley et al. 2012; Falush et al. 2003). This continual association has resulted in human and *H. pylori* genetic histories which broadly mirror each other, for example genetic distance increases and diversity decreases with increasing distance from Africa in both humans and *H. pylori* (Linz et al. 2007). *H. pylori* consists of several major genetic populations which correspond to large geographical areas: hpEurope, hpEastAsia, hpAsia2, hpNEAfrica, hpAfrica1 and hpAfrica2 (Falush et al. 2003; Moodley et al. 2012). Figure 4.1a shows the relationships between the major ancestral populations, and their extant counterparts are shown in Figure 4.1b.

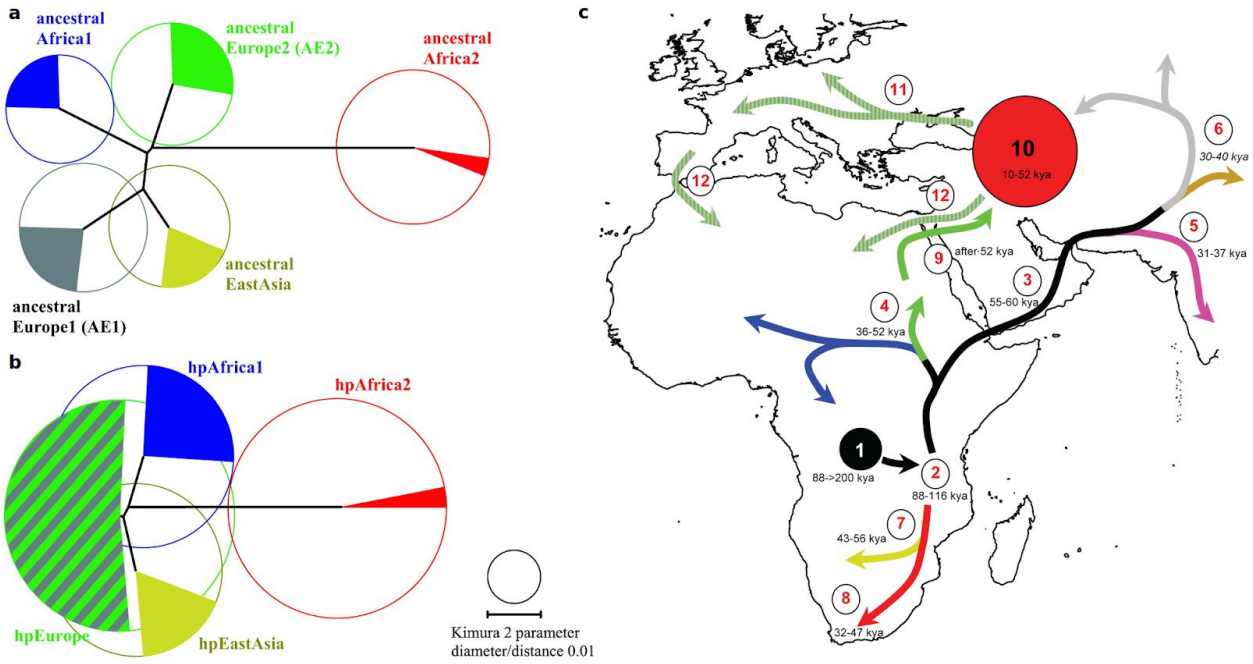

Figure 4.1: An overview of the major *H. pylori* populations and their origins. a. The five major ancestral *H.pylori* populations, and **b.**, their extant counterparts, with hpEurope shown as a hybrid between AE1 and AE2. The size of the circles corresponds to the within-population genetic diversity. The filled portions of the circles correspond to the numbers of isolates in the original datasets, and are irrelevant in this context. **c.** The current model for the evolutionary origins of the major *H. pylori* populations. The colours of both ancestral and extant hpAfrica2, hpAfrica1, hpEastAsia, and hpEurope are the same in all parts of the figure. Figure adapted from (Falush et al. 2003; Moodley et al. 2012).

The current model for how these populations emerged is shown in Figure 4.1c. *H. pylori* was acquired by humans at least 100,000 years ago (Figure 4.1c, 1) and differentiated into two major lineages (Figure 4.1c, 2), with one lineage evolving with the San people in South Africa, ultimately becoming hpAfrica2 (Figure 4.1c, 8). The other major lineage was carried out of Africa during the first successful out of Africa migration (Figure 4.1c, 3), and differentiated into hpAfrica1 and hpNEAfrica within Africa (Figure 4.1c, 4), and hpAsia2 and hpEastAsia in Asia (Figure 4.1c, 6) (Moodley et al. 2012). hpEurope was formed by introgression between two lineages: AE1 (ancestral Europe 1) from central and South-West Asia and AE2 (ancestral Europe 2) from North-East Africa, and is therefore a hybrid (Figure 4.1b, Figure 4.1c, 10). This introgression is thought to have started in the Middle East or Western Asia, and continued gradually North-West across Europe, where the hybrid strains replaced the ancestral European population (Moodley et al. 2012; Falush et al. 2003). This resulted in a cline, and the proportion of African ancestry decreases from the Southern hpEurope strains to their Northern counterparts.

This introgression was likely very disruptive, with extensive rearrangement of loci leading to many novel interactions throughout the genome. The recent advances in whole-genome sequencing mean there are now thousands of genome sequences for *H. pylori*, offering an opportunity to study evolutionary events in great detail. Here I present a study of the long-term consequences of the introgression in hpEurope, and show that selection has likely moderated the uptake of DNA from different ancestral sources.

Methods

Sequencing, mapping, and core genome definition

Genome assemblies were downloaded from BIGSdb, and synthetic sequencing reads were created from these assemblies using ArtificialFastqGenerator (Frampton and Houlston 2012). These synthetic reads were mapped to the 26695 *H. pylori* reference genome using Snippy (--mincov 10 --minfrac 0.9) (https://github.com/tseemann/snippy). [N.B. The synthetic read generation and mapping were performed by Kaisa Thorell, Karolinska Institutet, Stockholm, Sweden.] For each isolate, the '.consensus.subs.fa' file (containing the SNPs), and the '.aligned.fa' file (containing information on unmapped sites) were merged to create a consensus sequence containing both SNPs and information on unmapped sites. This step is important for creating a core genome, as when only the '.consensus.subs.fa' file is used the absence of SNPs within a genomic region can either mean that this region is identical to the reference, or that these sites are not present in the isolate. Distinguishing between these possibilities is important for estimating mutation rates. Genes and IGRs with > 90% sequence present in > 95% of isolates were used to create a core genome.

Calculation of dN/dS, dI/dS, and PSM

The pipeline used in chapter 1 was used to calculate these quantities.

Population structure analysis

Chromopainter and fineSTRUCTURE were used to assign individual isolates to populations (Lawson et al. 2012; Yahara et al. 2013). Chromopainter was first used, and for each region of DNA in each isolate, a likely donor was assigned from the other isolates in the dataset. This information was then used to produce a co-ancestry matrix showing the proportion of ancestry each isolate shares with every other isolate. This was then used as input to fineSTRUCTURE,

which classified the isolates into populations with distinct ancestry profiles. [N.B. The Chromopainter and fineSTRUCTURE analyses were performed by Koji Yahara, National Institute of Infectious Diseases, Tokyo, Japan, Kaisa Thorell, Karolinska Institutet, Stockholm, Sweden, and Daniel Falush, University of Bath, Bath, UK.]

Results

Selection on core genes and IGRs in *H. pylori* is comparable to other species

I first repeated the PSM and dN/dI/dS analyses from chapter 1 on 476 *H. pylori* isolates (Figure 4.2). In the PSM analysis, synonymous sites had the lowest PSM values (19%), followed by intergenic sites (36%), followed by non-synonymous sites (52%), and nonsense sites had the highest PSM values (75%) (Figure 4.2a). This pattern was the same when PDM values were calculated from doubleton mutations. These results are consistent with the results from other species, and suggest that selection on intergenic sites is intermediate between that acting on synonymous and non-synonymous sites. The dN/dS values range from 0.1-0.2, and the dI/dS values are significantly higher, ranging from 0.3-0.5 ($P < 10^{-15}$, Mann-Whitney *U* test) (Figure 4.2b). These results are consistent with those from other species, suggesting that selection is weaker on intergenic sites than on non-synonymous sites.

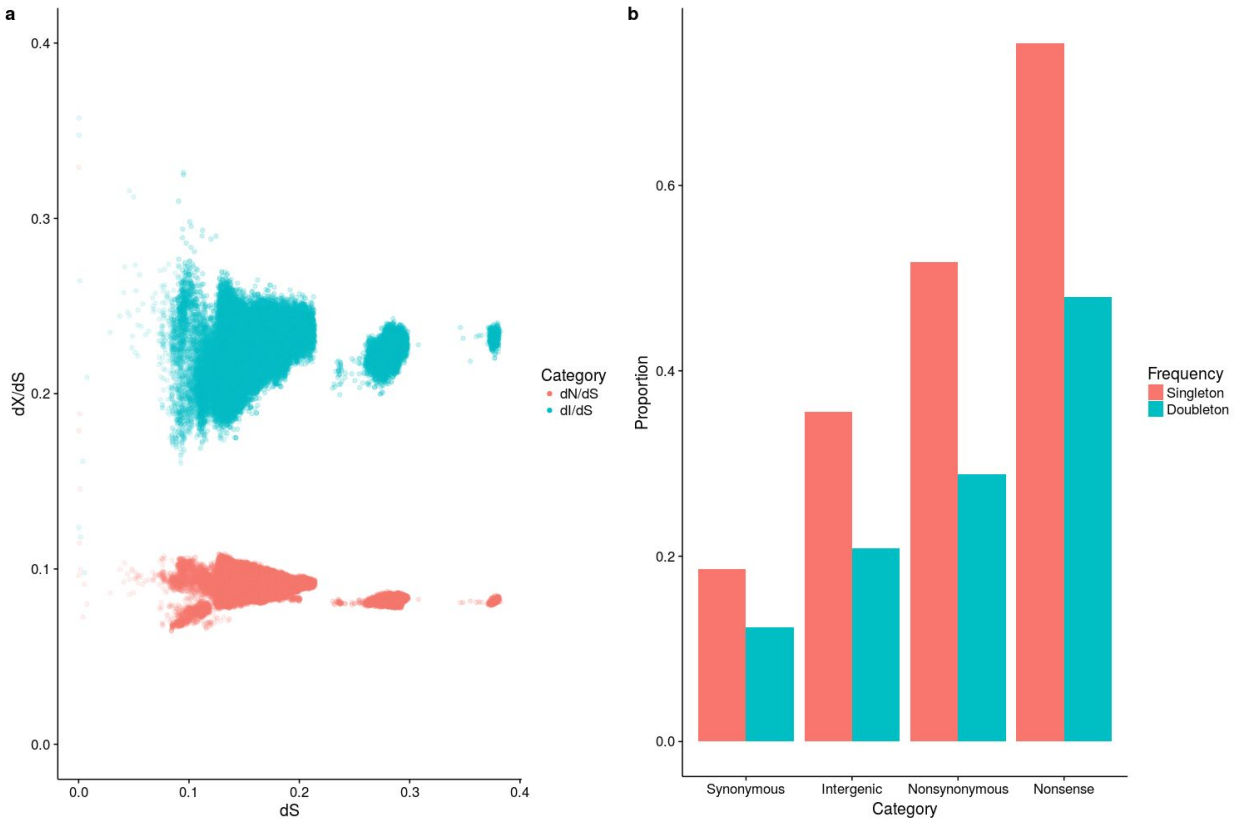

Figure 4.2: Purifying selection on *H. pylori* genomes. a. Pairwise dN/dS and dI/dS comparisons between 476 *H. pylori* isolates. dN/dS (and dI/dS) were plotted against dS as a measure of divergence time. **b.** Proportions of both singleton and doubleton mutations (those present in one or two isolates, respectively), in four different mutation categories.

Populations of *H. pylori* vary in their deleterious mutation load

H. pylori consists of several distinct genetic populations with different evolutionary histories. The African and Asian populations are ancient, distinct populations, and the European populations are hybrids of African and Asian ancestors (Falush et al. 2003; Moodley et al. 2012). To investigate the effect of these different evolutionary histories on the mutational load within contemporary populations, I calculated dN/dS values for each pairwise comparison between isolates, and separated the comparisons into within and between-population comparisons (Figure 4.3). For each comparison, I plotted dN/dS against dS to separate the populations from each other and to control for the decrease in dN/dS which has previously been reported (Rocha et al. 2006).

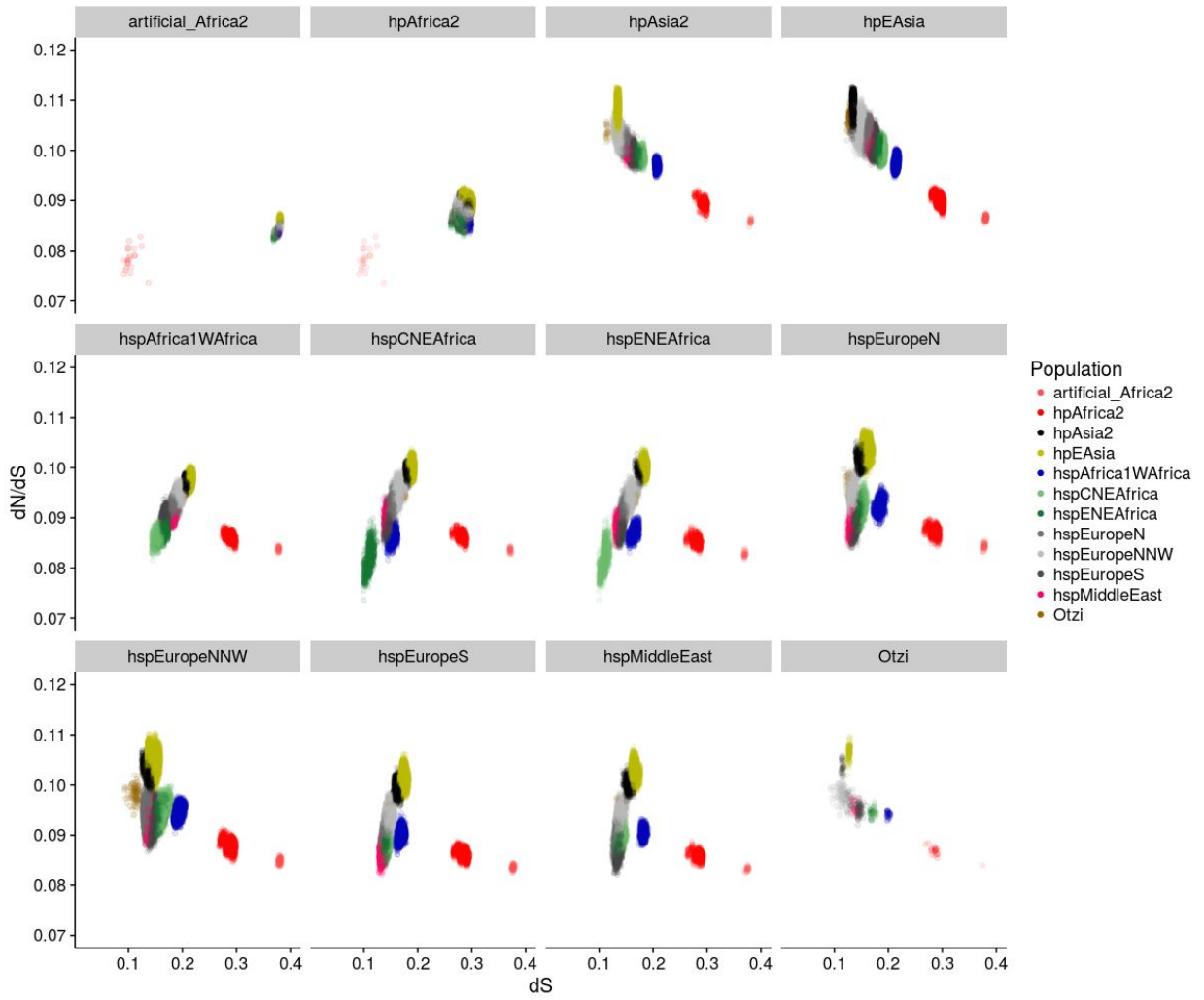

Figure 4.3: Purifying selection on *H. pylori* genomes, from between-population comparisons. Pairwise dN/dS comparisons between different *H. pylori* populations, where dN/dS is plotted against dS as a measure of divergence time. In each panel, there is a focal population, and within that panel each point is a comparison between an isolate from the focal population and an isolate from a different comparator population. The comparisons are coloured by the comparator population, and these colours are consistent with those in Figure 4.1.

In the between-population comparisons, there are clear and substantial differences in dN/dS values between populations, ranging from 0.07-0.11. In the Asian focal plots (hpEAsia and hpAsia2), the populations are separated by dS values, and the dN/dS values decrease with increasing dS. This means the Asian populations have the highest dN/dS values, followed by the European populations, and the African populations have the lowest dN/dS values. Given previous work, this could be interpreted as ongoing purifying selection acting to remove deleterious non-synonymous mutations over time (Rocha et al. 2006). However, the hspAfrica1 focal plots (hspAfrica1WAfrica, hspCNEAfrica, hspENEAfrica) show the opposite trend, where

dN/dS increases with increasing dS. This trend is however the same with respect to the populations, with gradually increasing dN/dS values from the African populations, to the European populations, to the Asian populations. This means that the difference between populations in dN/dS values must result from genuine differences between the populations in terms of population genetic processes, and not a difference in divergence time. The European populations are hybrids of African and Asian strains, and so in the European focal plots the populations are not well separated by dS. However, the Asian populations (hpAsia2, hpEAsia) have higher dN/dS values than the other populations, which is consistent with previous observations. I also plotted only within-population comparisons (Figure 4.4).

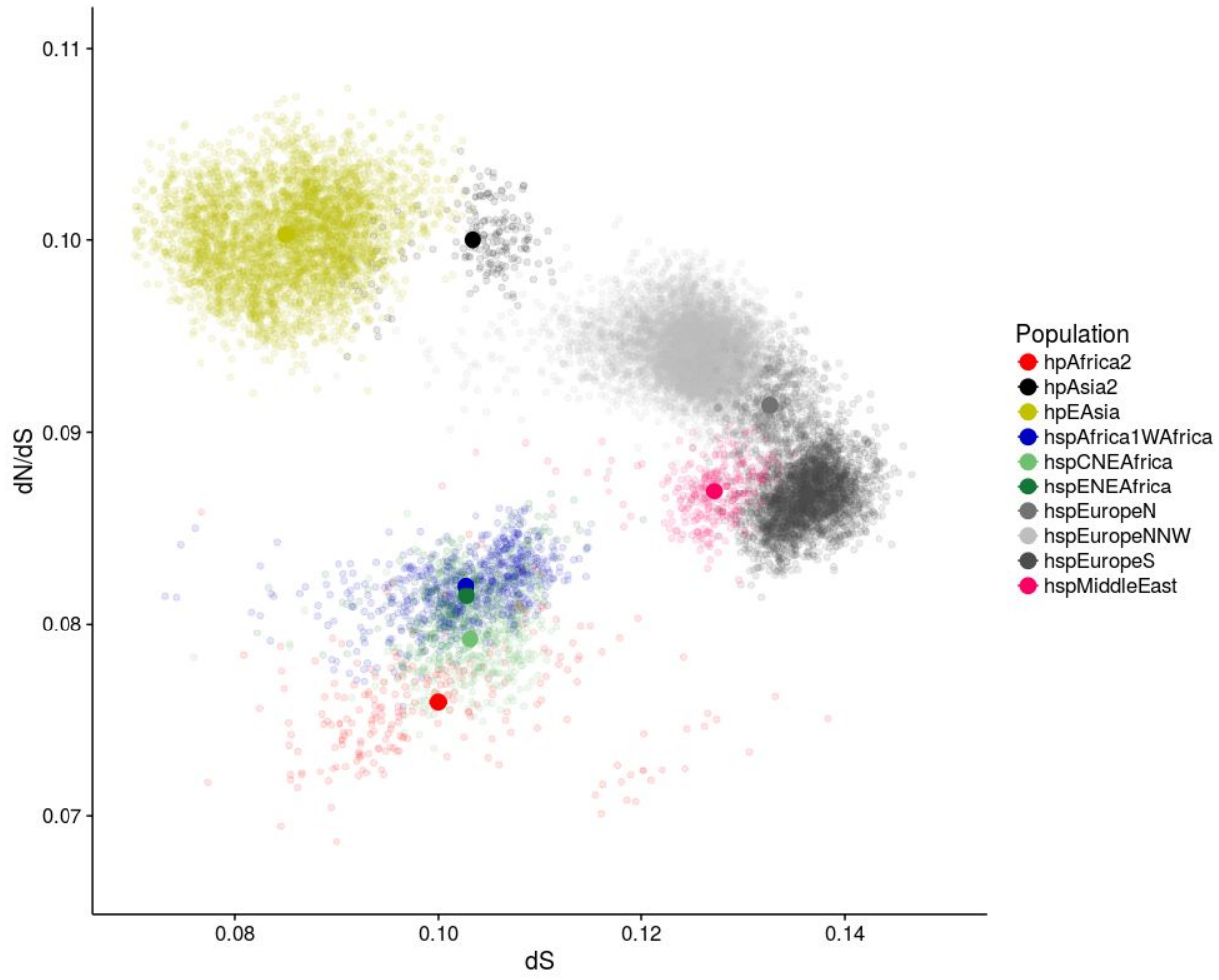

Figure 4.4: Purifying selection on *H. pylori* genomes, from within-population comparisons. Pairwise dN/dS comparisons between isolates from the same population, where dN/dS is plotted against dS as a measure of divergence time. The partially transparent circles are comparisons between individual isolates, and the filled circles with error bars correspond to the mean ± sem for each population. In some cases the error bars are too small to be seen. The colours are consistent with those in Figure 4.1.

In the within-population comparisons (Figure 4.4), there is again considerable variation in dN/dS values between the different populations. hpAfrica2 have the lowest dN/dS values, followed by hpAfrica1 populations, the European populations are intermediate, and the Asian populations have the highest dN/dS values.

In order to facilitate comparisons between the different populations, I used hpAfrica2 as an outgroup as it is divergent from the other populations. As introgression between hpAfrica2 and another population will affect (reduce) the genetic distance between the two populations, I constructed an artificial hpAfrica2 outgroup strain. I did this by painting the hpAfrica2 genomes

with fineSTRUCTURE, and choosing the least introgressed sites, thus creating an artificial hpAfrica2 strain with minimal introgression (hereafter referred to as hpArtAfrica2). [N.B. the fineSTRUCTURE analysis was performed by Koji Yahara, National Institute of Infectious Diseases, Tokyo, Japan.] The effect of this is shown in Figure 4.5, where the other populations are essentially equidistant from hpArtAfrica2 (as measured by dS). When compared with hpArtAfrica2, hpEAsia and hpAsia2 have the highest dN/dS values; this is consistent with previous observations. However, the African and European populations are less clearly distinguished from each other.

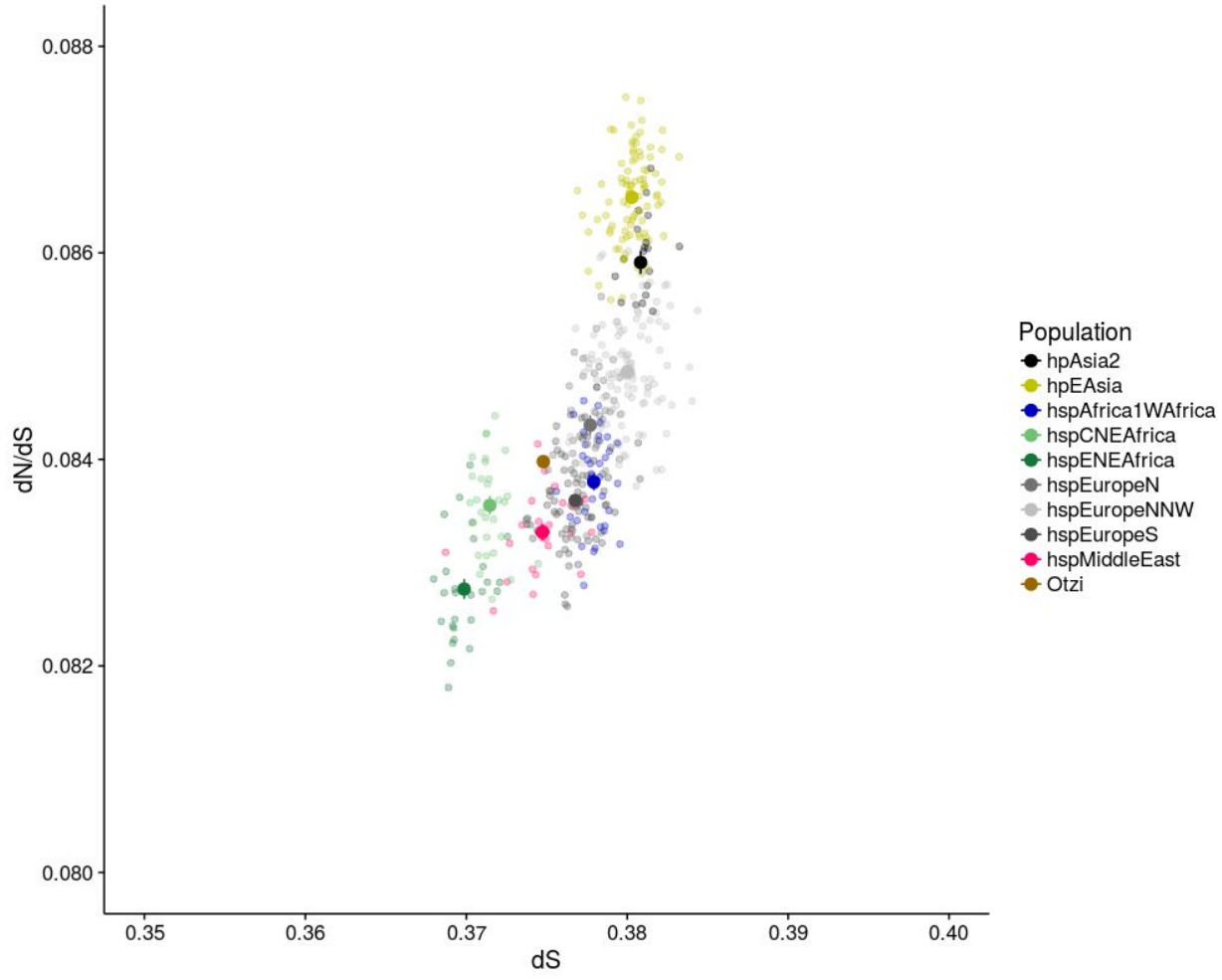

Figure 4.5: Purifying selection on *H. pylori* genomes, using hpArtAfrica2 as an outgroup. Pairwise dN/dS comparisons between hpArtAfrica2 and isolates from other populations, where dN/dS is plotted against dS as a measure of divergence time. The other populations are nearly equidistant from the outgroup, as shown by the limited differentiation of the populations by dS. The colours are consistent with those in Figure 4.1.

The introgression within European populations has been moulded by selection

To summarise the results so far, the two ancient separated populations from Africa and Asia have low and high dN/dS values, respectively. The European populations are hybrids between African and Asian strains, and have intermediate dN/dS values. As the European populations have had access to both African and Asian DNA, it is possible that their genetic composition is a result of selection for the fittest DNA from either source. To test this possibility, I combined information from estimates of ancestry with information about mutation frequencies. I analysed three European populations (hspEuropeS, hspEuropeN, hspEuropeNNW), and for each focal European population I painted the genome with hpEAsia and hspAfrica1WAfrica as source populations. [N.B. The chromosome painting was performed by Koji Yahara, National Institute of Infectious Diseases, Tokyo, Japan.] This provided site-by-site estimates for the ancestry of the European populations. I then calculated a mutation score for the hpEAsia and hspAfrica1WAfrica source populations, where each population was separately compared to the hpArtAfrica2 outgroup. For each for each site I calculated a score from 0-1 based on the frequency of mutations within a source population compared to hpArtAfrica2. The score was defined as number of isolates with the mutation divided by the number of isolates within the source population, so a site with a fixed mutation in the source population has a score of 1, and a site with no mutation within the source population has a score of 0. These scores were calculated separately for each source population, and separately for non-synonymous and synonymous mutations. For each mutation type, I then calculated the difference between the scores for the two source populations at each site, to give a score between -1 (hpEAsia specific mutations) and 1 (hspAfrica1WAfrica specific mutations). It must be noted here that mutations which are specific to either source population are likely to be present in the European populations as a result of introgression. Thus, for each site I have calculated both a mutation score (for both non-synonymous and synonymous mutations), and an estimate of ancestry. I then calculated the mean of these quantities for each gene, and this data is shown in Figure 4.6.

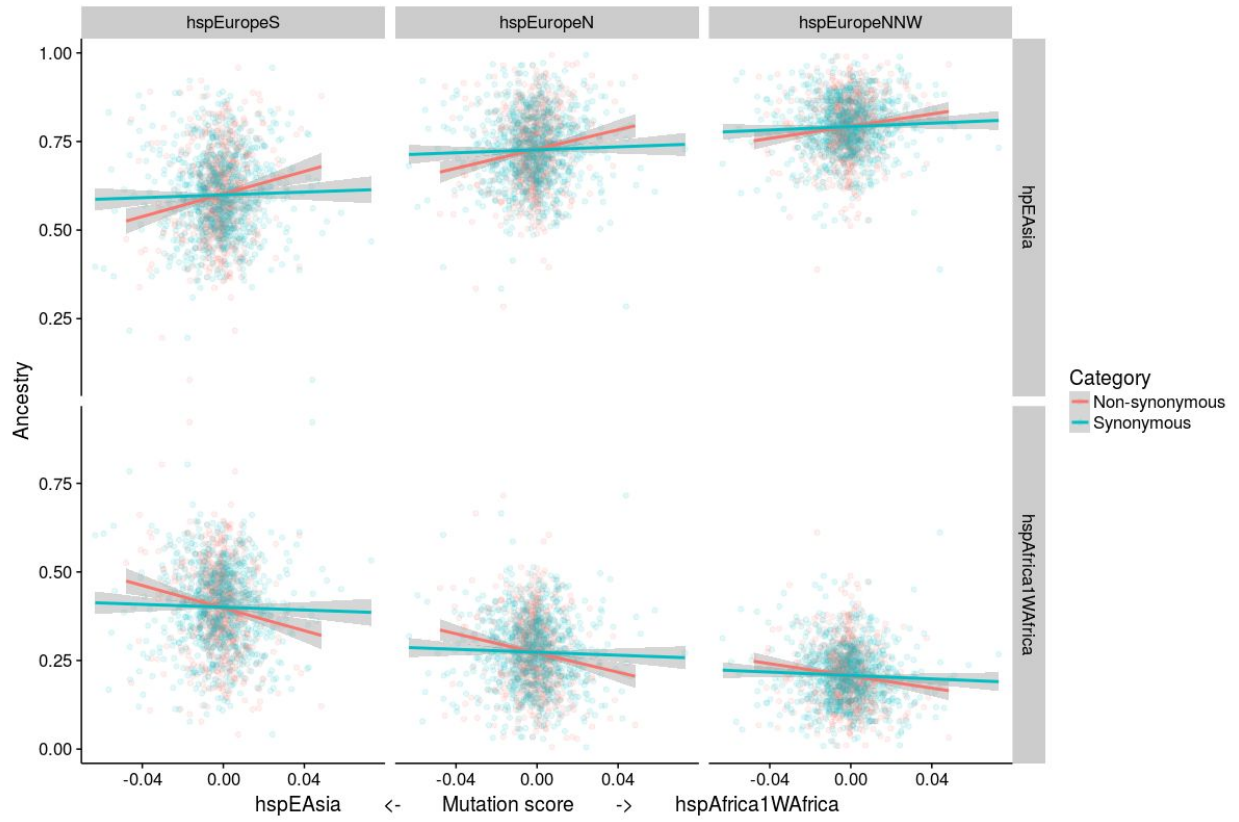

Figure 4.6: The source of population-specific mutations predicts ancestry in hpEurope populations. Two donor populations (hpEAsia and hspAfrica1WAfrica) were chosen as proxy source populations for hpEurope. Within each of these populations, a score was calculated for each mutation based on its frequency within each source population. These scores were then combined with ancestry estimates for the hpEurope populations. The mean scores and ancestry estimates for each gene were calculated (for both synonymous and non-synonymous sites), and these are shown (each point is a gene).

For non-synonymous mutations in each focal European population, the mutation score predicts the ancestry (Figure 4.6). In genes with strongly negative scores, (those with many mutations present between hpEAsia and the outgroup which are not present in hpAfrica1WAfrica), the hpEAsia ancestry component is lower, and vice-versa. Crucially, this effect is much stronger for non-synonymous mutations than for synonymous mutations, meaning that non-synonymous mutations within source populations can better predict ancestry than synonymous mutations. This suggests that selection is likely to explain the difference between non-synonymous and synonymous mutations.

There are also differences between the three hpEurope populations. The proportion of African ancestry decreases moving South-North, and the strength of selection on the introgression also

decreases from hspEuropeS to hspEuropeNNW (as shown by the steepness and differentiation of the non-synonymous and synonymous regressions, Figure 4.6). This is consistent with a model of introgression which started in the Middle East or Western Asia, and continued North-West across Europe, where donor DNA gradually became more limited (Falush et al. 2003; Moodley et al. 2012).

This analysis relies on the assumptions that European isolates are perfect hybrids of Asian and African strains, and that Africa2 is an appropriate outgroup (i.e. it is approximately equidistant from the African and Asian populations at every gene). The former is supported by the literature; when painted with the major old-world populations, the European populations overwhelmingly take their ancestry from hpAfrica1WAfrica and hpEAsia, even when hpAfrica2 is included as a donor (Falush et al. 2003; Linz et al. 2007; Thorell et al. 2016). The latter is supported by a further analysis of the mutation data (Figure 4.7). For each gene, the synonymous distance to the outgroup was calculated for hspAfrica1WAfrica and hpEAsia. This analysis confirmed that these two populations are almost exactly equidistant from the outgroup (hpAfrica2), and that there are no genes which show evidence of recent admixure with the outgroup (as these would have extremely low synonymous distances to the outgroup). This means that for each gene, the nucleotides present in a European population could reasonably have come from either an Asian or African source. It also assumes that the introgression which gave rise to the European populations has happened recently, and that much of the source DNA present at the time is still present in modern descendents of AE1 and AE2. Both of these assumptions are supported by recent evidence, which suggests that the introgression may have happened as recently as 6000 years ago (Maixner et al. 2016; Moodley et al. 2012).

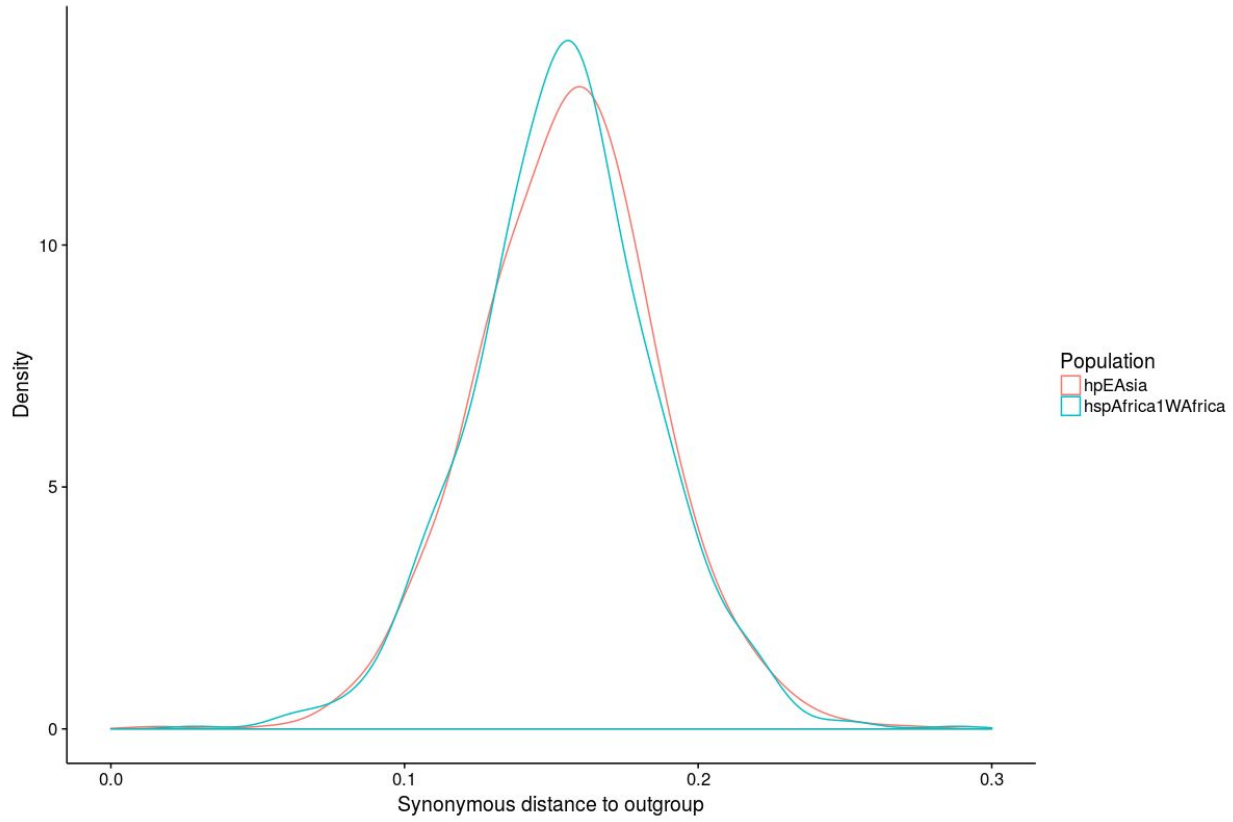

Figure 4.7: The synonymous distances between the source populations and the outgroup. The synonymous distance was calculated for each gene between each source population and the outgroup. The synonymous distance is the sum of the synonymous scores across the gene divided by the length of the gene.

Discussion

The separation of *H. pylori* into differentiated populations which largely mirror human populations has been well studied. These populations correspond to large geographical areas, and contain enough evidence to reconstruct major events (such as the out of Africa migrations) (Falush et al. 2003; Moodley et al. 2012). These populations have likely been subject to varying levels of selection and drift as a result of their different histories. From analysis of dN/dS (and dI/dS) within and between populations, I have noted that there is substantial variation in these quantities, with hpAfrica2 having the lowest dN/dS values. This population is thought to have been associated with the San people in South Africa for at least 100,000 years, and limited movement of this population may have prevented bottlenecks, resulting in very strong purifying selection. The West African populations had slightly higher dN/dS values. The East Asian populations had the highest dN/dS values, and this is consistent with reduced efficacy of purifying selection, resulting from a reduced effective population size due to bottlenecks. It has

previously been shown that as distance from Africa increases, the diversity of the populations decreases. This is consistent with my results, which show that the efficacy of purifying selection decreases with distance from Africa, presumably as a result of reduced diversity (reduced effective population size).

The exception to this model is hpEurope, which is thought to have been formed by gradual introgression between EA1 (of Asian descent), and EA2 (of African descent) (Falush et al. 2003; Moodley et al. 2012). This was likely very disruptive, as the mixing of Asian and African DNA will have introduced many novel combinations of loci, bringing both deleterious and beneficial interactions. Here, I have attempted understand the long-term effects on the genome of the introgression. I have done this by characterising regions of the genome with a preponderance of mutations in proxies for AE1 (hpEAsia), and EA2 (hspAfrica1WAfrica), compared with an outgroup (hpArtAfrica2). I find that in genes with many hpEAsia specific mutations (relative to the outgroup and hspAfrica1WAfrica), the ancestry of these genes within European populations is more likely to be from hspAfrica1WAfrica than hpEAsia. This effect is much stronger for non-synonymous compared to synonymous mutations, and the effect is present when the populations are reversed (i.e. genes with many hspAfrica1WAfrica specific mutations are more likely to have ancestry from hpEAsia). This is not a result of limited genetic divergence between the outgroup and the source populations, for example by recent admixture. To summarise, when a gene has many non-synonymous mutations specific to one source population, in European populations it is preferentially sourced from the other source population. A compelling explanation for the observed difference in this effect between non-synonymous and synonymous mutations is that selection has acted to limit the uptake of population specific mutations into the European strains.

Chapter 5

Compensatory evolution is widespread in Rho-independent terminators in bacteria

Introduction

Terminating transcription is an important part of bacterial gene regulation, and it is achieved by two primary mechanisms: Rho-dependent termination, and Rho-independent (intrinsic) termination (Peters, Vangeloff, and Landick 2011; Santangelo and Artsimovitch 2011). Rho-dependent termination utilises Rho, along with a number of auxiliary proteins to achieve transcription termination. There are few consistent motifs present at Rho-dependent terminators, so they cannot be accurately predicted from the genome sequence (Ciampi 2006). In contrast, Rho-independent termination relies on the intrinsic properties of the terminator mRNA sequence. These consist of a GC-rich stem loop structure followed by a U-rich tract, and these properties mean they can be accurately predicted from the genome sequence (Peters, Vangeloff, and Landick 2011; De Hoors et al. 2005; Kingsford, Ayanbule, and Salzberg 2007). Rho-independent termination occurs as follows: The U-tract induces a pause in transcription, the hairpin then nucleates to form the stem-loop structure, as the hairpin nucleates, this pulls the DNA-RNA hybrid apart, and RNA polymerase dissociates (Peters, Vangeloff, and Landick 2011). The two methods of termination are used to different extents by different bacterial species. In the gram-positive low-GC firmicutes, Rho-independent transcription is the dominant method, and in *Staphylococcus aureus* Rho is not essential (Washburn et al. 2001; Kingsford, Ayanbule, and Salzberg 2007). In contrast, *E. coli* uses the two methods more equally, and Rho is an essential gene in this species (Ciampi 2006).

The intrinsic sequence properties of Rho-independent terminators mean that they can be accurately predicted from the genome sequence (De Hoors et al. 2005; Kingsford, Ayanbule, and Salzberg 2007), and these properties also enable the evolutionary dynamics of such sequences to be studied. A recent study showed that Rho-independent terminators are subject to purifying selection in bacteria (Thorpe et al. 2017). Further, the stem was shown to be under stronger purifying selection than the loop. This is expected as mutations within the stem are likely to disrupt the complementary Watson-Crick base pairing, and these are likely to be deleterious. In contrast, mutations within the loop are likely to be more neutral as they are not constrained in the same way as the stem.

Mutations within the stem which disrupt the complementary Watson-Crick base pairing may be compensated by a second mutation which repairs the complementary base pairing (illustrated in Figure 5.1). This phenomena was recently shown to be common in *Bacillus cereus*, suggesting

that strong selection acts to maintain these sequences (Safina, Mironov, and Bazykin 2017). Here I provide a complementary analysis of Rho-independent terminators in a number of bacterial species, and find that compensatory evolution is widespread in bacteria.

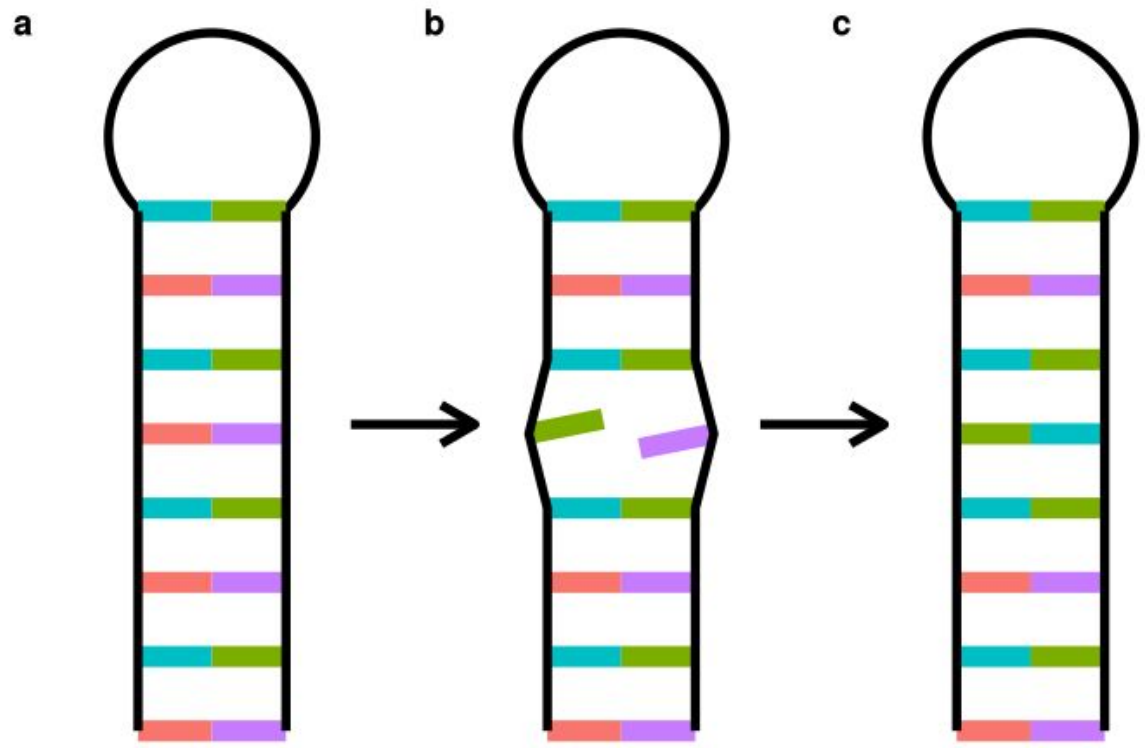

Figure 5.1: Compensatory evolution in a stem-loop sequence. a. An intact stem-loop sequence. **b.** A mutation occurs on the left side of the stem, disrupting the complementary Watson-Crick base pairing. **c.** A second mutation occurs on the right side of the stem, repairing the stem.

Methods

Dataset

For *E. coli*, *S. enterica*, *S. pneumoniae*, *K. pneumoniae*, *S. aureus*, and *M. tuberculosis*, the datasets used in chapter 1 were used. For *H. pylori*, the dataset used in chapter 3 was used. For *R. salmoninarum, V. anguillarum,* and *V. parahaemolyticus*, the isolates were obtained from collaborators. For each species, overrepresented clonally related isolates (such as those belonging to the same sequence type) were removed to leave a dataset where each major lineage was represented, but only by a limited number of isolates.

Sequencing, mapping, and SNP calling

The isolates were mapped against a single reference genome for each species using SMALT-0.7.6 (https://sourceforge.net/projects/smalt). SAMtools-0.1.19 (Li et al. 2009) was used to produce Variant Call Format (VCF) files, which were filtered to call SNPs. SNPs were only called if they passed all of the following thresholds: depth >= 4, depth per strand >= 2, proportion of reads supporting the SNP >= 0.75, base quality >= 50, map quality >= 30, af1 >= 0.95, strand bias >= 0.001, map bias >= 0.001, tail bias >= 0.001. Consensus Fasta sequences were then used to produce an alignment for each species. For *V. anguillarum* and *V. parahaemolyticus*, the two chromosomes were treated separately. [N.B. The mapping for the *R. salmoninarum*, *V. anguillarum*, and *V. parahaemolyticus* data was performed by Nicola Coyle and Sion Bayliss, University of Bath, Bath, UK.]

Genome annotation and core genome definition

Each reference genome was annotated using Prokka-1.11 (Seemann 2014). The terminator predictions were produced using TransTermHP (Kingsford, Ayanbule, and Salzberg 2007), and obtained from the PePPER webserver (de Jong et al. 2012). These annotations were used to extract genes and terminator sequences, and a core genome was produced consisting of genes and terminators with > 90% sequence present in > 95% of isolates.

Calculation of dN/dS and dI/dS

The pipeline used in chapter 1 was used to calculate these quantities.

Identification of compensatory mutations

I identified compensatory mutations in a pairwise manner between isolates. In a comparison between two isolates, if a SNP was present at the complementary position of both the left and right hand side of the terminator stem, this was identified as a potential compensation event. These events were then classified as 'compensatory' if the nucleotides on each side of the stem formed a Watson-Crick pair in both isolates, and 'non-compensatory' if they did not. Thus, for a compensatory mutation to be identified, there must be a SNP in both the left and right sides of the stem, and in each isolate the nucleotides at these positions must be complementary (Figure 5.1).

Compensatory mutation simulation

In order to provide a neutral expectation to compare against (as compensatory mutations would be expected to occur by chance), I performed simulations. I simulated mutations on terminator sequences and counted the number of compensatory mutations observed. Mutations were simulated with both no mutation bias, and the observed mutation bias where GC -> AT mutations are more common than the reverse (Hershberg and Petrov 2010; Hildebrand, Meyer, and Eyre-Walker 2010). Using singleton mutations (those present in only one genome) I calculated the per-site relative mutation rates for the 6 mutation types for each species. As these biases were similar in all species, I averaged them across all species and used this as the observed mutation bias.

Results

Properties of terminators

I assembled datasets of 10 diverse bacterial species, of which 6 corresponded to those in a previous study of selection on intergenic sites (*E. coli, S. enterica, K. pneumoniae, S. pneumoniae, S. aureus,* and *M. tuberculosis*) (Thorpe et al. 2017). I supplemented these with *V. anguillarum, V. parahaemolyticus, R. salmoninarum,* and *H. pylori*. The two vibrio species both have two chromosomes, and so offered the opportunity to study differing selective constraints on the different chromosomes (Cooper et al. 2010). *R. salmoninarum* is a high GC low-diversity species, with similar genomic properties to *M. tuberculosis*. *H. pylori* was included because there is some uncertainty in the literature about the extent to which it uses Rho-independent termination (De Hoors et al. 2005; Washio, Sasayama, and Tomita 1998; Castillo et al. 2008).

Terminators were predicted using TransTermHP (Kingsford, Ayanbule, and Salzberg 2007). Although the properties of Rho-independent terminators enable them to be predicted from the sequence, this does not guarantee that they are functional. The algorithm used by TransTermHP accounts for the presence of both the GC-rich stem-loop and U-tail in order to minimise the erroneous prediction of sequences which contain only one of these features (for example uptake sequences which consist of a GC-rich stem-loop). The firmicutes (*S. aureus* and *S. pneumoniae*) are known to rely predominantly on Rho-independent termination (Rho is not essential in *S. aureus*). Additionally, the accuracy of prediction was shown to be extremely high in the closely related species *Bacillus subtilis*, where 88% of experimentally confirmed terminators were predicted with a 2.1% false positive rate (De Hoors et al. 2005; Kingsford, Ayanbule, and Salzberg 2007). Other species with many high quality terminator predictions

were members of the *Vibrio* genus, although there is no experimental data to confirm these. In *M. tuberculosis*, predicted terminators were shown to be functional by analysis of RNA-seq data (Botella et al. 2017). There are conflicting reports of the extent that *H. pylori* uses Rho-independent termination, but one study provided experimental evidence that terminators predicted by TransTermHP were functional (De Hoors et al. 2005; Washio, Sasayama, and Tomita 1998; Castillo et al. 2008). Although these studies provide good evidence that the predictions contain functional terminators, without further experimental data this cannot be confirmed, and the analysis must be considered with this caveat.

I first analysed the general properties of terminator sequences (Figure 5.2). Terminator stems had mean lengths of approximately 6 bp in most species, but varied from 6 bp in *H. pylori* to 10 bp in *S. aureus*. The loops were more consistent, with mean lengths of 4-5 bp. These figures are in agreement with previous research (Peters, Vangeloff, and Landick 2011; Kingsford, Ayanbule, and Salzberg 2007). The stems had higher GC contents than the loops, and intriguingly the GC content of the stems decreased from the foot of the stem to the loop ($P < 0.05$, Spearman's correlation) for all species except *H. pylori* and *K. pneumoniae*. This observation of decreasing GC content is novel, and may be explained by selection for stronger GC bonds towards the foot of the stem.

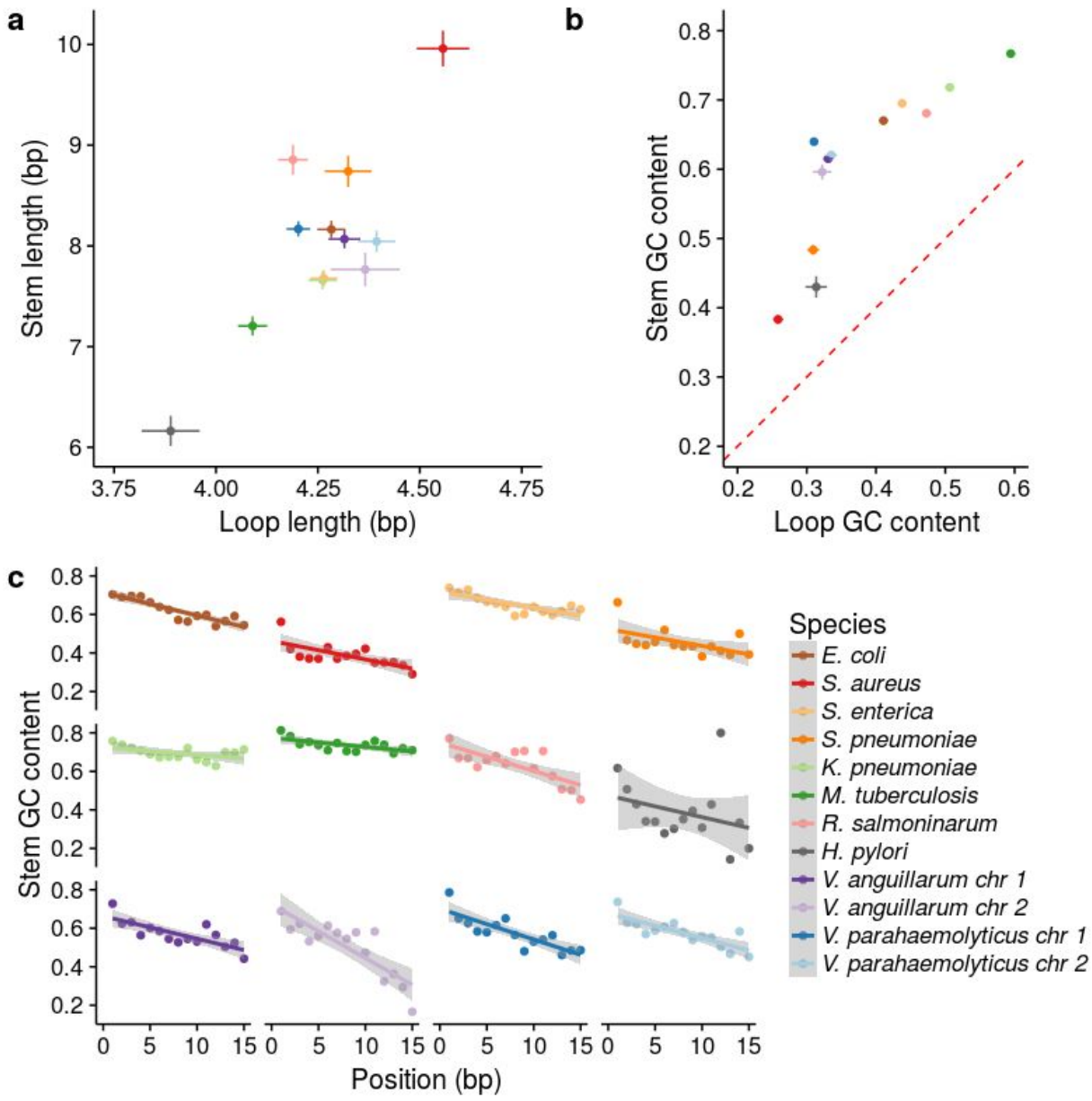

Figure 5.2: Properties of Rho-independent terminators. a. Terminator stem and loop lengths. The points and error bars correspond to the mean ± sem. **b.** Terminator stem and loop GC contents. The points and error bars correspond to the mean ± sem. **c.** Stem GC contents against position in the stem, from the foot end (0) to the loop end (15). Stems were limited to 15 bp in length as there were very few longer than this. In all cases the two chromosomes of *V. anguillarum* and *V. parahaemolyticus* were analysed separately.

Purifying selection on terminators

I used dI/dS (a modification of dN/dS to work on intergenic sites) to investigate purifying selection on terminators (Figure 5.3a). I analysed the stem and loop separately, as in (Thorpe et al. 2017). As in (Thorpe et al. 2017), I find that both the stem and loop parts of terminators are

subject to purifying selection in most species, and that the stems are consistently more constrained than the loops ($P < 10^{-15}$, Mann-Whitney *U* test) in all species except *R. salmoninarum, M. tuberculosis,* and *H. pylori*, where there is no significant difference between the stem and loop parts. *R. salmoninarum* and *M. tuberculosis* display wide variation in dI/dS values, typical of recently differentiated strains, or more generally low-diversity species (Rocha et al. 2006; Castillo-Ramírez et al. 2011; Thorpe et al. 2017). *V. anguillarum* and *V. parahaemolyticus* offer an opportunity to compare selection pressures on different chromosomes, as both species have two chromosomes. In both species, dI/dS from the terminator stems (and dN/dS from the core genes) were lower in chromosome 1 than 2 ($P < 10^{-15}$, Mann-Whitney *U* test). This is consistent with previous work showing that purifying selection is stronger on chromosome 1 than 2 (Cooper et al. 2010).

I also analysed the locations of mutations within the terminator stems, as in (Safina, Mironov, and Bazykin 2017). In agreement with Safina *et al.*, I find that mutations are significantly more likely to be observed at the first and last (external) positions of the stem than those within the stem (internal) ($P < 0.05$, Fisher's exact test). This was true for all species except *R. salmoninarum, M. tuberculosis, H. pylori* and chromosome 2 from *V. anguillarum* (Figure 5.3b). I also tested for differences between the number of mutations at the first (foot) and last (loop) stem positions, as done by Sefina *et al*. I found only limited evidence for this, with only *K. pneumoniae*, *V. anguillarum* chromosome 1, and *V. parahaemolyticus* chromosome 1 reporting significant results ($P < 0.05$, Fisher's exact test) (Figure 5.3c).

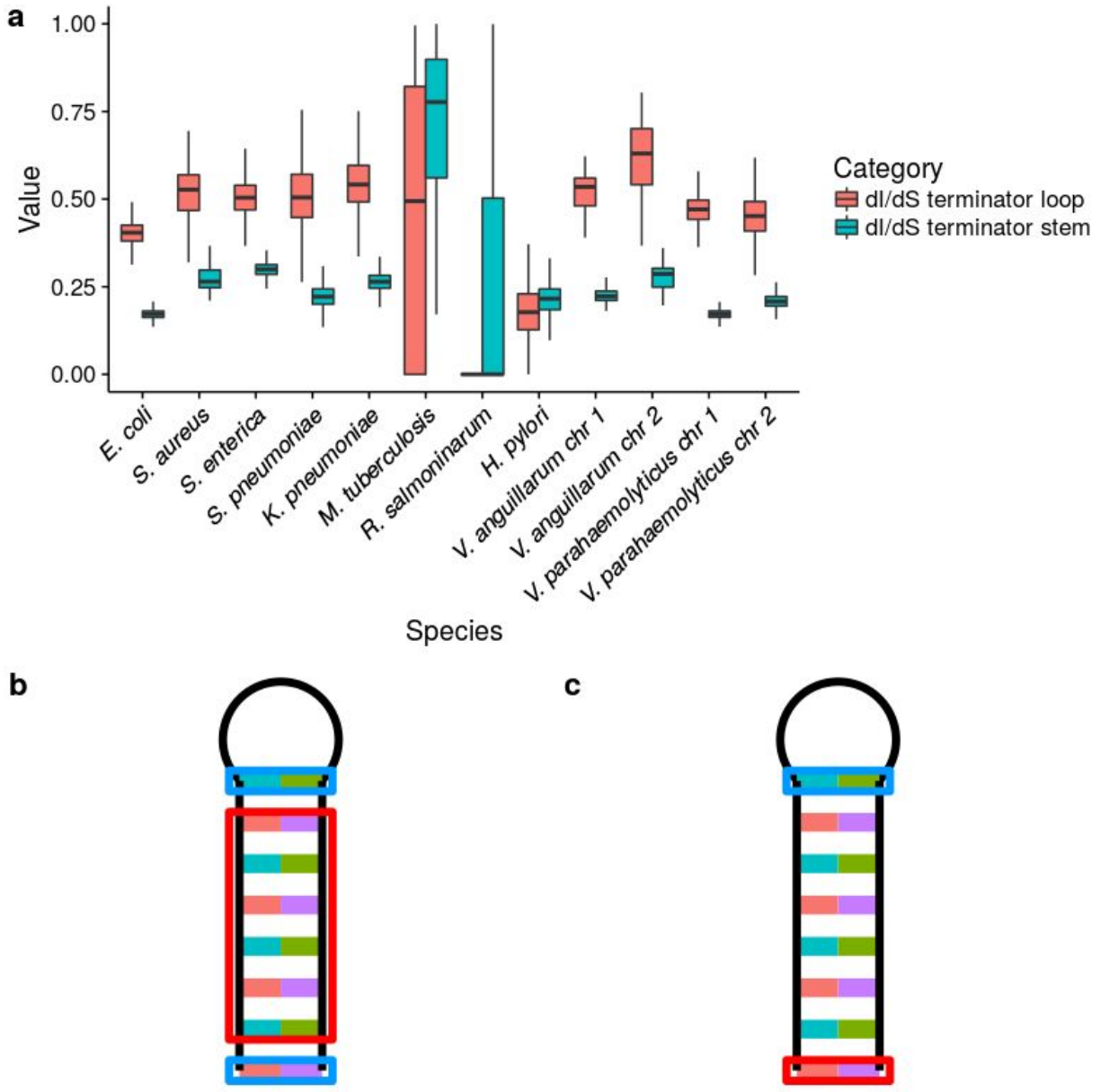

Figure 5.3: Purifying selection on Rho-independent terminators. a. dI/dS values for terminator stem and loops. The two chromosomes of *V. anguillarum* and *V. parahaemolyticus* were analysed separately. **b.** Comparisons between the numbers of mutations per site in external (blue) and internal (red) positions within the stem. **c.** Comparisons between the numbers of mutations per site in the external positions within the stem, at the loop end (blue), and foot end (red). In both **b** and **c** the blue positions are those with an excess of mutations, and this colouring is consistent with (Safina, Mironov, and Bazykin 2017).

Compensatory evolution is widespread

In a recently published study, (Safina, Mironov, and Bazykin 2017) showed that compensatory evolution in terminator sequences was relatively common in *Bacillus cereus*. I used a slightly

different method to identify compensatory mutations in my datasets. Whereas Safina *et al.* performed ancestral state reconstruction, I analysed isolates in a pairwise manner. I did this because some of the species analysed (such as *V. parahaemolyticus* and *H. pylori*) recombine at extremely high rates, resulting in star-like phylogenies with low bootstrap support values at the base of the deep branches (Cui et al. 2015). Where mutations were present on both sides of complementary positions within the terminator stem, these were identified as potential compensatory events. If in both strains being compared, these positions formed a Watson-Crick complementary pair, then they were classified as 'compensatory', and if not then they were classified as 'non-compensatory'. I also simulated mutations on terminator stems to provide a neutral expectation for the rate of compensatory evolution, and these simulations were performed both with and without mutation bias.

This analysis revealed that compensatory mutations are widespread across bacteria (Figure 5.4). For all species except *H. pylori, R. salmoninarum, M. tuberculosis,* and the second chromosome on the two vibrio species, the number of compensatory mutations observed was significantly greater than the neutral expectation ($P < 0.05$, Mann-Whitney *U* test). These results were robust to simulations performed both with and without mutation biases. Further, the rates of compensatory mutations were higher than those of non-compensatory mutations (where mutations are present at both positions but do not form a Watson-Crick pair). These results indicate that compensatory mutations are strongly selected for. The rate of compensatory evolution was highest in *E. coli*, this is likely a product of the large effective population size of this species (Charlesworth 2009). In *V. anguillarum* and *V. parahaemolyticus*, the rate of compensation was higher on the first chromosome compared to the second (where there was no evidence of higher rates compared to the neutral expectation). This suggests that selection is stronger on the first chromosome, in agreement with the dN/dS values reported above (Cooper et al. 2010). *H. pylori* shows no evidence of selected compensatory evolution. I tested for positional biases of compensatory mutations (as in Figure 5.3b), but found no evidence of this.

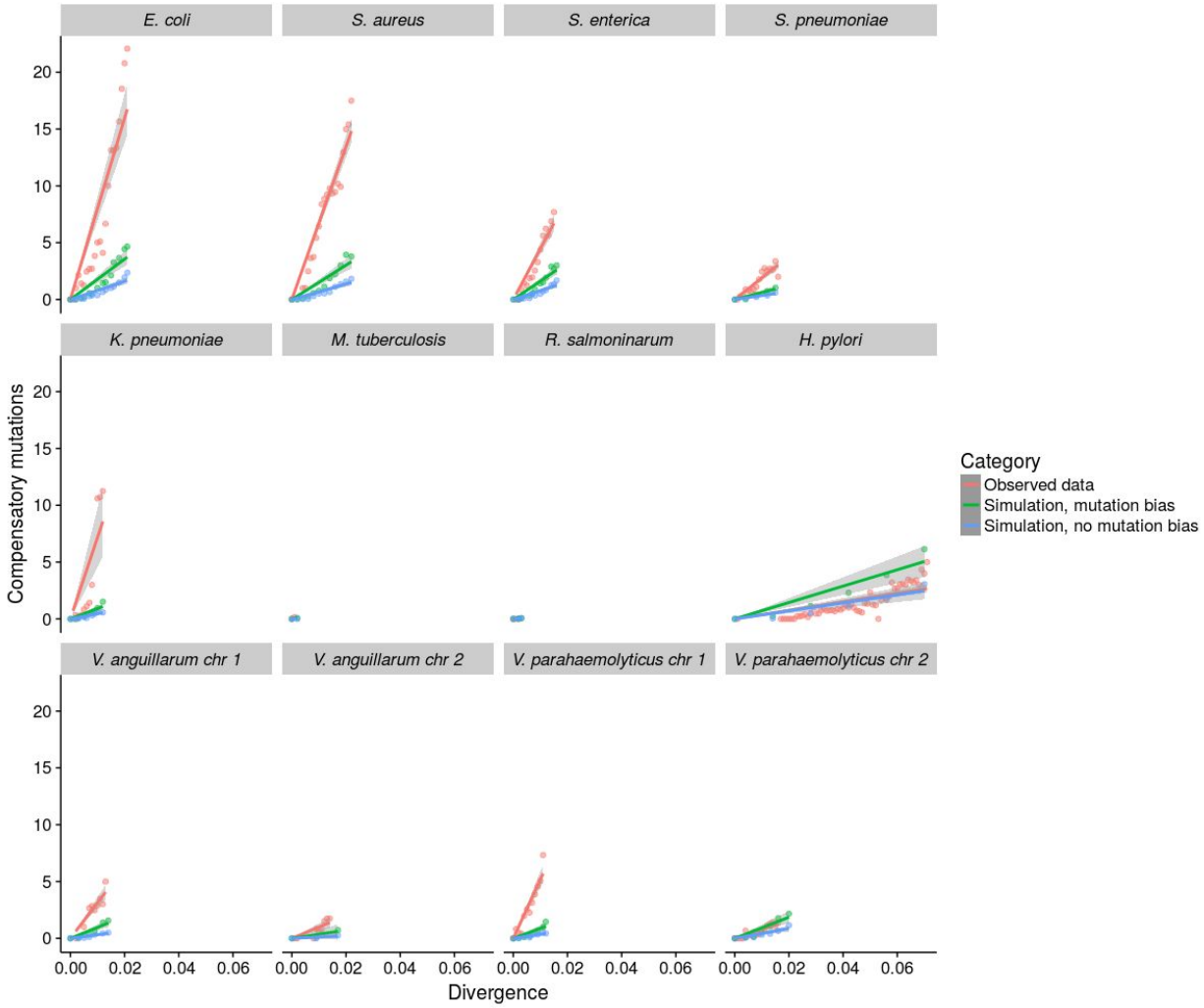

Figure 5.4: Compensatory evolution of terminator sequences. The number of compensatory mutations was plotted against the divergence within terminator sequences. Each panel corresponds to a different species, and the two chromosomes of *V. anguillarum* and *V. parahaemolyticus* were analysed separately. The red points correspond to the observed number of compensatory mutations, and the green and blue points correspond to the simulated data.

Discussion

Here I have performed a detailed analysis of the evolutionary dynamics of Rho-independent terminator sequences in bacterial genomes. Consistent with Safina *et al.*, I have shown that terminator sequences are under strong purifying selection, which increases in strength from the loop, to the external stem positions, to the internal stem positions. I also find that the strength of selection varies according to species. *H. pylori*, which does not rely heavily on intrinsic termination (Kingsford, Ayanbule, and Salzberg 2007), shows weak evidence of selection, whereas *E. coli* (which utilises intrinsic terminators and has a large N_e) shows strong evidence of selection. In the species with two chromosomes (*V. anguillarum* and *V. parahaemolyticus*),

the terminators on the first chromosome are under stronger selection than those on the second chromosome. This is consistent with previous work showing that genes on the first chromosome are more strongly selected than those on the second chromosome.

I also provide an analysis of compensatory evolution which is both complementary to, and broader than that performed previously (Safina, Mironov, and Bazykin 2017). My approach enabled me to perform simulations to provide a neutral expectation for the rate of compensatory evolution, and to record the rate of non-compensatory mutations. This analysis revealed widespread selected compensatory evolution, providing further evidence that these sequences are subject to strong selective pressures. Signals of selection from individual species were consistent with previous observations; *E. coli* had high rates of compensatory evolution, compared to the low rates observed in *H. pylori*, and in the two *Vibrio* species the first chromosome exhibited higher rates of compensatory evolution than the second.

The explanations for high rates of compensatory evolution are not immediately clear. Strong selection is presumably required to explain the number of observed compensatory mutations, however this strong selection would also be expected to select against any initial mutations which disrupt the stem-loop structure, eliminating them from the population by purifying selection. Perhaps the terminators which are compensated are not the most strongly selected, as it is only in these terminators that the intermediate state (one mutation) is tolerated. The strength of selection on a terminator is likely to be governed by the expression of its upstream gene, as the cost of inefficient termination will be greater in highly expressed genes. Another factor which is likely to influence the strength of selection on terminators is changes in effective population size (N_e). Deleterious mutations (such as a single mutation in a terminator) can rise in frequency during bottlenecks, and perhaps these are compensated when N_e increases and selection is more efficient. Investigating both the between-terminator variation and the effect of N_e on compensatory evolution are areas for further research, with the caveat that N_e is a difficult quantity to measure in bacteria (Charlesworth 2009; Cui et al. 2015; Price and Arkin 2015; Didelot et al. 2011; Sharp et al. 2005; Hartl, Moriyama, and Sawyer 1994).

R. salmoninarum and *M. tuberculosis* displayed very little evidence of selection, with widely varying dI/dS values and no enrichment of compensatory mutations. This may be because Rho-independent termination is not important in these species (Washio, Sasayama, and Tomita

1998), or that selection is genuinely weak. Previous work on *M. tuberculosis* has consistently reported high dN/dS values of approximately 0.6, indicating that purifying selection is weak in this species (Hershberg et al. 2008; Thorpe et al. 2017). However, there are reports of positive and diversifying selection in *M. tuberculosis* (Pepperell et al. 2013; Farhat et al. 2013; Osório et al. 2013). *H. pylori* also showed limited evidence of selection on terminator sequences, with no difference between the stem and loop dI/dS values and no enrichment of compensatory mutations. In contrast with the two low-diversity species discussed above, *H. pylori* is a diverse species with genome-wide dN/dS values of approximately 0.09. Thus, *H. pylori* is subject to strong purifying selection, and the lack of evidence for selection on terminators may be because the sequences are not functional in this species. This is consistent with previous observations (De Hoors et al. 2005; Washio, Sasayama, and Tomita 1998), but in contrast with more recent work (Castillo et al. 2008).

In conclusion, for most species there is widespread evidence of both purifying selection, and compensatory evolution in Rho-independent terminator sequences. In contrast, *H. pylori* shows little evidence of selection on terminator sequences, despite strong genome-wide purifying selection, indicating that terminator sequences are not important in this species.

Chapter 6

Discussion

Discussion

The work presented in this thesis is broad in scope, and has tried to advance our understanding of genome evolution in bacteria, with a focus on intergenic sites. Although the precise questions and organisms under study are different in each chapter, they are complementary to each other and form a coherent set of analyses. Chapter 1 is concerned with measuring the selective constraint on intergenic sites in the core genomes of a diverse set of bacterial species. Chapter 2 is also concerned with intergenic sites, but here both core and accessory sites are considered, and a new tool, Piggy is introduced to facilitate these analyses. As in chapter 1, chapter 3 is an analysis of selection, but in this chapter *Helicobacter pylori* is studied in detail to understand how extensive admixture may be moderated by selection. Finally, chapter 4 is focused on compensatory evolution in transcriptional terminator sequences. A common thread throughout the thesis is an attempt to understand the diversity present within bacterial genomes, and how these genomes are shaped by selection, with particular focus to those regions of the genome which are poorly understood on a genomic scale, such as intergenic sites. Here I discuss the major results and themes of the chapters, and place them into the context of our current understanding. Finally, some directions for further research are considered.

Bacterial genomes are strongly shaped by selection

Previous work has shown that mutations which are likely to have very subtle fitness effects are influenced by selection in bacteria. This includes selection on codon bias, AT skew, and genomic GC content (Sharp et al. 2005; Charneski et al. 2011; Hildebrand, Meyer, and Eyre-Walker 2010). That these features are selected in bacteria shows that selection must be a strong force, and this can be explained by large population sizes and short generation times in many bacterial species (*E. coli* is estimated to have a long-term effective population size of 25,000,000 (Charlesworth 2009)). Since large whole-genome datasets consisting of hundreds of isolates from a given species have become available, there has been much effort to study the protein-coding components of bacterial genomes, but comparatively little effort to study their intergenic components.

One small study of group A Streptococcus genomes, a study of Buchnera genomes, and a study of 22 diverse bacterial clades all found evidence of purifying selection on intergenic sites (Luo et al. 2011; Degnan, Ochman, and Moran 2011; Molina and Van Nimwegen 2008). The work presented in chapter 1 provides both complementary and additional analyses to this

literature. The overall observation that intergenic sites are subject to purifying selection is consistent with previous work. The PSM analysis in chapter 1 is similar to that done by (Luo et al. 2011) as they are both based on the site-frequency spectrum, and the calculation of dI/dS (as a modification of dN/dS) is similar in principle to the calculation of *R* values in (Molina and Van Nimwegen 2008). Despite differences in approach, these analyses, along with the Kmer based analysis in (Degnan, Ochman, and Moran 2011), all show intergenic sites to be under purifying selection which is intermediate in strength between that acting on synonymous and non-synonymous sites. Additionally, in chapter 1 different regulatory elements were analysed separately, and ribosome binding sites and non-coding RNAs were found to be under stronger purifying selection than the other elements, and the terminator stems were more constrained than the loops. These novel analyses advance our understanding by showing that not all intergenic sites are equally important, and that I can measure differences in constraint between these classes of site.

Selection moderates the effect of introgression

In chapter 3, a detailed analysis of introgression in *H. pylori* was carried out. It was shown that in European populations (which are a hybrid between Asian and African bacteria), the source of introgressed DNA was likely to have been shaped by selection. The DNA of the European populations was painted with proxies for the two source populations in order to identify likely ancestry, and this was then compared to mutation data. Areas of the European genomes with many Asian specific mutations were more likely to be inherited from African strains and vice-versa, and this effect was stronger for non-synonymous mutations than synonymous mutations. This suggests that selection has moderated the introgressed DNA to reduce the load of deleterious non-synonymous mutations.

This work builds on previous research showing differences in mutational load in other species. In *S. aureus*, recombined fragments often have lower dN/dS values than mutation-derived SNPs when very closely related isolates (such as those belonging to the same clonal complex) are compared (Castillo-Ramírez et al. 2011). This is likely explained by the age of these mutations; mutations in the recombined fragments are likely to be considerably older than recent mutation-derived SNPs. As selection has had longer to purge older deleterious mutations, this results in lower dN/dS values in the recombined fragments. This effect of time dependence on dN/dS is explained in detail in (Rocha et al. 2006), and this provides an important consideration

when comparing dN/dS values. Additionally, dN/dS values have been shown to vary in response to ecological shifts, such as the higher dN/dS values observed in intracellular *Shigella* species compared to free-living *E. coli* (Balbi, Rocha, and Feil 2009). These studies show that inferences of selection vary according to the age of mutations, regions of the genome, and ecological conditions, and the novel work on *H. pylori* adds to this by showing that selection moderates deleterious mutation load according to the source of introgressed DNA.

Compensatory evolution is widespread

The work presented in chapter 4 provides a more detailed analysis of Rho-independent terminator sequences in a range of bacterial species. This follows recent work by (Safina, Mironov, and Bazykin 2017) which showed that compensatory evolution is relatively common in *Bacillus anthracis*. Chapter 4 shows that compensatory evolution is widespread among bacteria, suggesting that these elements are subject to strong selection. The rates of compensation are also consistent with expectation in some species, for example *E. coli* (with a large N_e) has high compensation rates, whereas *H. pylori* (which does not rely heavily on Rho-independent termination) has low rates. Additionally, in the species with two chromosomes (*V. anguillarum* and *V. parahaemolyticus*), compensation rates are higher in the first chromosome than the second. This is consistent with previous work showing that the first chromosome is subject to stronger selection than the second (Cooper et al. 2010). The study by (Safina, Mironov, and Bazykin 2017) was the first to show compensatory evolution in terminator sequences in bacteria. However, this study only investigated one species (*B. anthracis*), and so general statements could not be drawn. The work in chapter 4 broadened the study of compensatory evolution, showing that it is widespread, but also showing variation between species which is consistent with previous knowledge and expectations.

Pan-genomes incorporating intergenic regions

The work discussed thus far clearly demonstrates that bacterial genomes are strongly shaped by selection, and that this selection is measurable on, and differs between classes of intergenic site. However, this work is based only on the core genomes of bacteria. The work in chapter 2 aimed to broaden our understanding of bacterial evolution by incorporating intergenic sites into pan-genome analyses. In order to facilitate this, a tool, Piggy, was developed to analyse IGRs in conjunction with Roary (Page et al. 2015). Chapter 2 showed that it is possible to create pan-genomes from IGRs in much the same way as from protein-coding sequences. There is

sufficient conservation within IGRs that a core set of IGRs are present, but there is more diversity in IGRs than genes overall. This is consistent with work in chapter 1 showing that IGRs are subject to weaker purifying selection than genes.

As gene regulation depends on elements located within IGRs, changes in IGRs can influence gene expression, and this can have profound phenotypic consequences. A striking recent example is a C -> T SNP in the -10 region of the promoter of *ptgE* from African *Salmonella enterica* serovar Typhimurium (Hammarlöf et al. 2017). This SNP resulted in a tenfold increase in the expression of PtgE, an outer membrane protein. This resulted in increased survival in human serum, and increased cleavage of Complement Protein B in vivo. Further, the SNP was required for successful infection in a chicken model. Another prominent example is a SNP in the promoter of the *eis* gene in *M. tuberculosis*; this SNP results in increased expression of Eis, and this confers resistance to kanamycin (Casali et al. 2012).

Although these are likely extreme examples, the work presented in chapter 1 shows that intergenic sites *en masse* are subject to purifying selection, and so many other SNPs within regulatory elements are likely to have more subtle fitness effects. One way to study this is to identify genes present in multiple isolates from a species, where that gene is preceded by divergent IGR sequences in different isolates. If the upstream IGR sequences are important for controlling gene expression then these genes would be expected to be differentially expressed in isolates with divergent IGRs. This was previously shown to be the case in *E. coli*, where up to 12% of the overall variance in gene expression could be attributed to this phenomenon (Oren et al. 2014). In the work presented in chapter 2, four *S. aureus* clonal complexes were analysed, and genomic and RNA-seq data were combined to investigate the same effect. In 9/12 comparisons, genes with divergent IGRs were more differentially expressed than those without, in accordance with the previous work on *E. coli*. From these results I can conclude that changes in IGRs frequently affect gene expression, and this is not limited to rare high-effect mutations. As gene expression in bacteria is tightly controlled (and therefore presumably under selection), this provides an explanation for the selection observed on IGRs in chapter 1.

Why do bacteria have such large pan-genomes?

It is clear from previous work that many bacterial species have enormous pan-genomes consisting of many thousands of genes (McInerney, McNally, and O'Connell 2017; Andreani,

Hesse, and Vos 2017; Holt et al. 2015; McNally et al. 2016), and from the work in chapter 2 that this is also true of IGRs. Advances in whole-genome sequencing, along with tools such as Roary and Piggy, have greatly increased our understanding of the scale of bacterial pan-genome variation (what is there), but this does not necessarily inform on deeper evolutionary questions (why is it there?). Assessing the evolutionary dynamics of these large pan-genomes is a clear direction for future research, and this will likely involve incorporation of population genetic ideas and methods with large-scale genomic data.

Several recent studies have started to address these questions, with surprisingly different results. (Andreani, Hesse, and Vos 2017) showed that pan-genome fluidity (a measure of pan-genome size), correlates with synonymous diversity across a number of bacterial species. According to population genetic models (Kimura 1991), species with large N_e maintain more neutral genetic diversity. Further to this, a number of recent studies have shown that rates of HGT are higher than point mutation rates in many bacterial species (Vos et al. 2015). From these findings, I could conclude that pan-genomes in bacteria are largely neutral and are governed by the effective population size of the species. The high rates of HGT would provide rapid turnover of genes, and providing that these genes are not overly deleterious, they would be maintained within a large pan-genome. (McInerney, McNally, and O'Connell 2017) argue that this scenario is unlikely because maintaining unnecessary genes is likely to be costly, and purifying selection would act to remove these genes. They use *E. coli* as a prime example, with a large estimated long-term N_e of 25 million, and ample evidence of purifying selection elsewhere in the genome (for example codon bias) (Charlesworth 2009; Sharp et al. 2005). They argue the large pan-genome in *E. coli* is likely due to migration into new niches, and this is consistent with mathematical models (Niehus et al. 2015). In support of this, pan-genome sizes correlate with lifestyles, with generalist environmental species having large pan-genomes, and host restricted endosymbionts having small pan-genomes (McInerney, McNally, and O'Connell 2017).

There is likely truth to both of these theories, and the dynamics of pan-genomes are likely governed by several competing forces. Perhaps one under appreciated component of these dynamics is time. (Rocha et al. 2006) showed that it is critical to incorporate time into analyses of selection, as deleterious mutations are gradually removed from populations over time, consistent with the nearly-neutral theory of evolution (Ohta 1973). For example, when a sample

of extremely closely related isolates from a bacterial species are analysed (such as those from the same sequence type), there is a preponderance of deleterious variation compared to when more distant isolates are compared (Rocha et al. 2006). This manifests as high dN/dS values, compared to low values observed between distant isolates. It is curious that one of the pan-genome examples used in (McInerney, McNally, and O'Connell 2017) consisted of 288 *E. coli* ST131 isolates. These isolates are extremely closely related (they are members of the same sequence type), and so are likely to contain large amounts of transient deleterious variation. A comparison between the mutational variation and gene content variation would shed much light on the evolutionary dynamics of the pan-genome here. For example, if the site frequency spectrum of genes and mutations was compared, this would enable patterns of transient gene content variation to be elucidated. Many singleton genes (and correspondingly mutations), would strongly indicate a situation where these genes are likely to be lost over time (as is observed in mutations). Alternatively, clusters of genes present in distinct groups of isolates would indicate some form of niche adaptation, and this would be more consistent with the hypothesis proposed by (McInerney, McNally, and O'Connell 2017).

Genotype-phenotype relationships

Understanding how genotypes contribute to phenotypes is one of the central goals of biology. Recent advances in GWAS methodologies have helped to identify mutations which contribute to phenotypes, but these tend to be limited to mutations with high penetration (Lees et al. 2016). Additional complications for bacteria arise from the clonal frame; many mutations are co-inherited along with causal mutations, and this limits the power of such approaches (Lees et al. 2016; Earle et al. 2016; Sheppard et al. 2013). A recent study by (Galardini et al. 2017) combined genomic and experimental data to predict phenotypes in *E. coli*. Genomic data from 696 strains was used to build loss-of-function probabilities for each gene, and this was integrated with conditional essentiality data from the *E. coli* K-12 reference strain. Using this information, accurate phenotype predictions were made for 38% of 214 tested conditions. This indicates that combining experimental and genomic data can provide great insight into genotype-phenotype causality. However, this study did not include information on intergenic sites, and so gene regulation was not considered within the analysis. It is possible that for some genes, their regulation is physiologically important, rather than simply the presence of the gene, or mutations contained within a gene. One such example is the promoter SNP in *S. enterica*

discussed above (Hammarlöf et al. 2017). A clear direction for improvement of this approach would be to incorporate information on gene regulation into the model.

However, the relationship between the divergence of an IGR and the expression of its cognate genes is not well understood. For example, a high effect mutation within a promoter may have more impact than several mutations in less important parts of the IGR. One approach could be to combine RNA-seq and genomic data to identify IGR variants which are associated with high and low expression of their cognate genes (as was done in chapter 2). This would then enable genes to be scored according to expression, and this would enable expression to be incorporated into the analysis.

Concluding remarks

The work presented in this thesis has shown that bacterial genomes evolve under strong selective constraints, and these constraints act upon intergenic sites, particularly the elements responsible for gene regulation. Where extensive introgression has taken place, selection has acted to moderate the deleterious mutation load by preferentially selecting loci from certain ancestry sources. These findings add weight to the view that there are no truly neutral sites in bacteria (Rocha and Feil 2010). The work on pan-genomes suggests that bacterial genomes should be viewed in terms of rich interactions between genes and IGRs, instead of only their gene content. Future work should focus on the interplay between these factors within appropriate evolutionary contexts to further our understanding of these important organisms.

Appendix

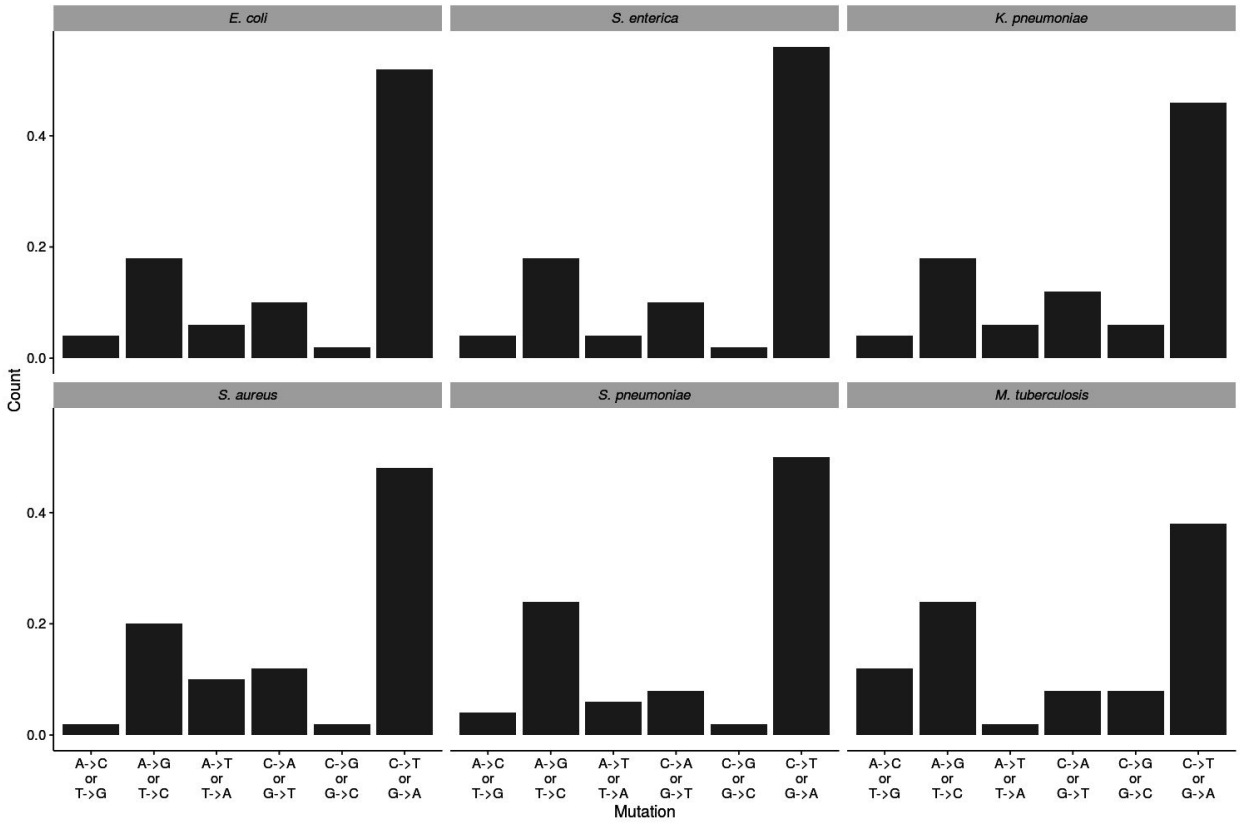

Figure S2.1: Per site mutation biases. The per site mutation bias was calculated for each species from the singleton SNPs. This was done by dividing the number of SNPs by the number of sites (e.g. the number of A->C and T->G mutations was divided by the number of A and T sites in the genome). These rates were then converted to a proportion so that the mutation types summed to 1 for each species.

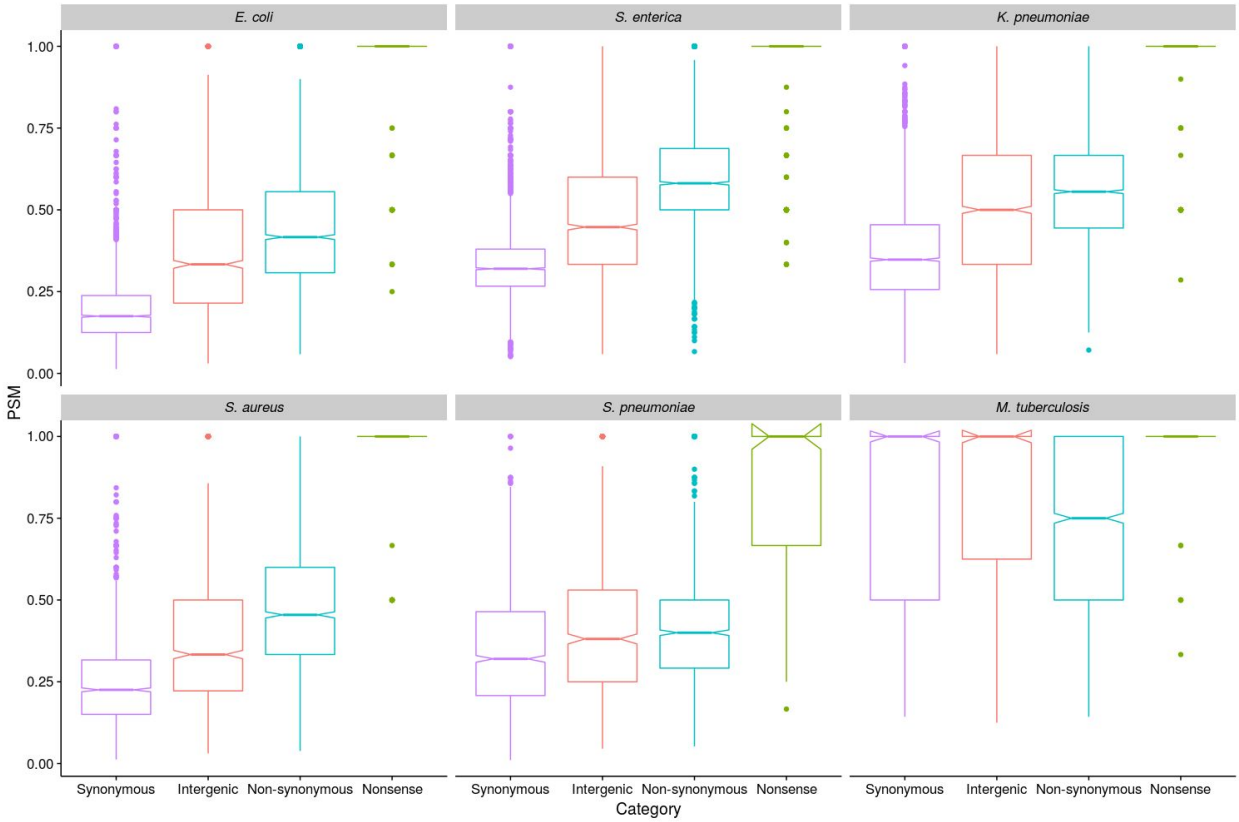

Figure S2.2: Analysis of selection on individual genes and IGRs. PSM values were calculated for each gene and IGR separately to check that our analysis was not confounded by a small number of highly conserved, unrepresentative IGRs. The notches in the box plots represent 95% confidence intervals around the median.

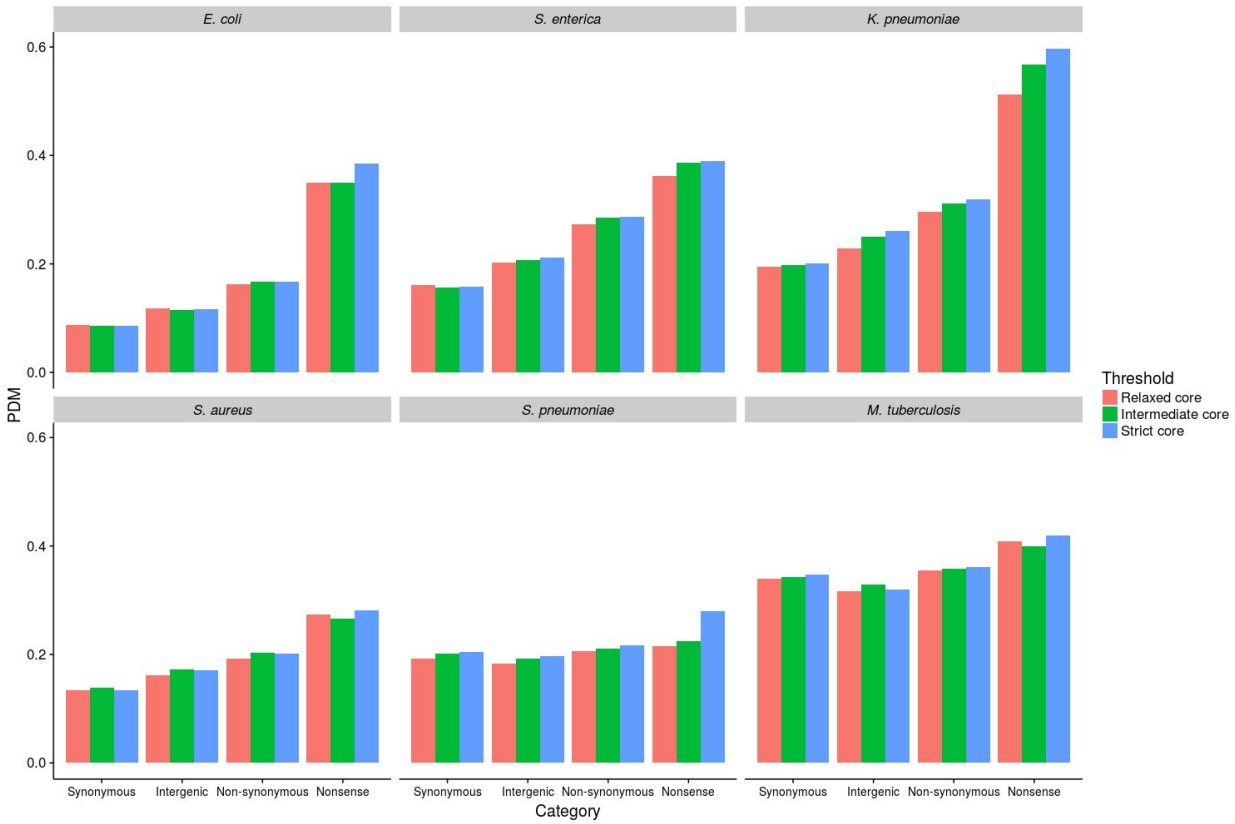

Figure S2.3: PDM (Proportion of Doubleton Mutations) analysis of selection on different mutation categories. PDM values were calculated by dividing the number of doubleton SNPs (those present in two genomes) by the total number of SNPs within that mutation category (excluding singletons).

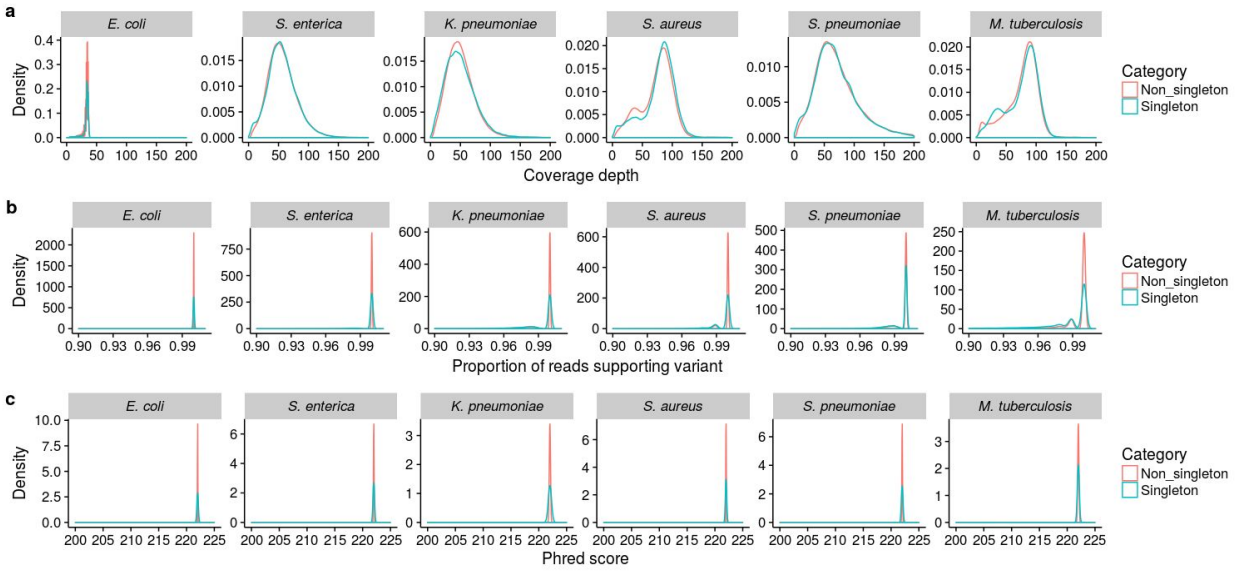

Figure S2.4: Validation of the quality of singleton SNPs. Singleton and non-singleton SNPs were analysed in order to validate the quality of the singleton SNPs. **a.** Depth of coverage of SNP positions. **b.** The proportion of reads supporting the SNP. **c.** The Phred Quality Q score of SNP positions.

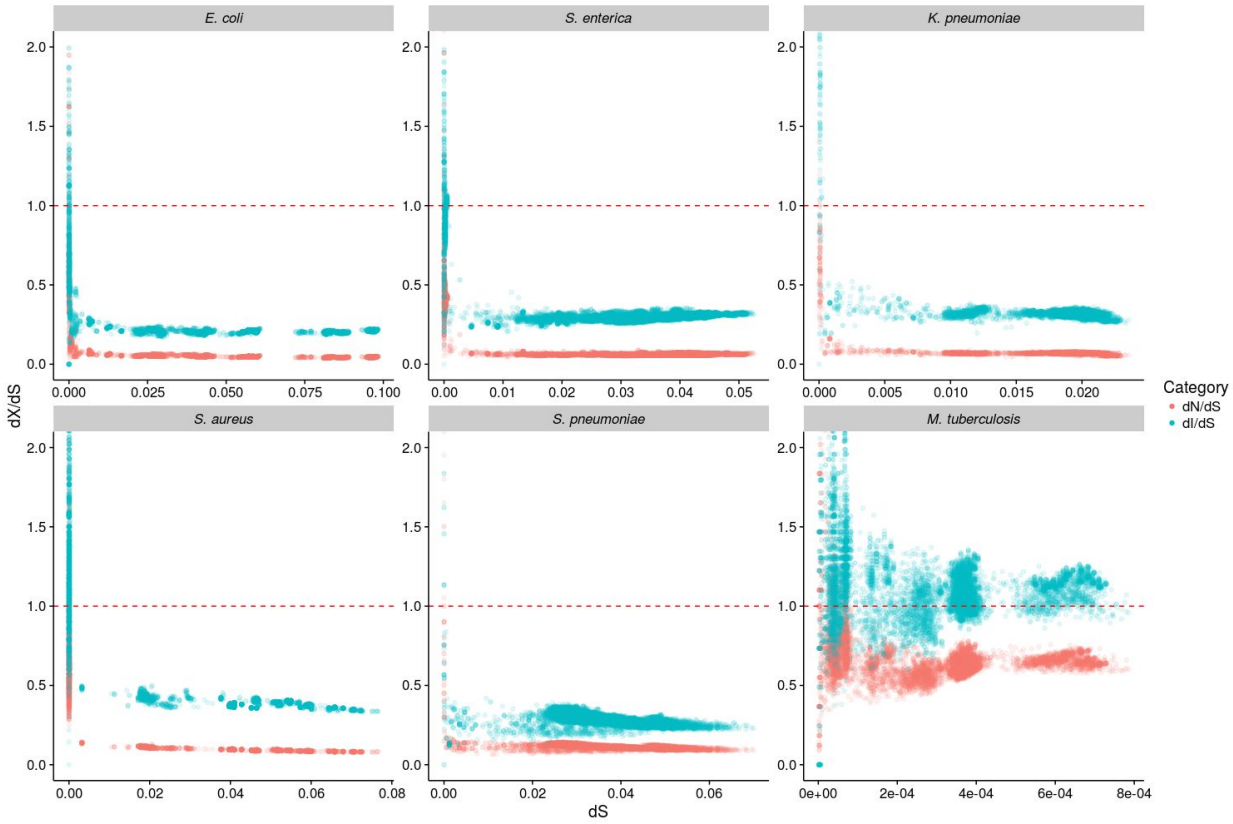

Figure S2.5: dN/dS and dI/dS analysis of selection. dN/dS and dI/dS were calculated between isolates in a pairwise manner, and these were both plotted against dS to explore the effect of divergence time on observed levels of selection. In order to control for the non-independence between the axes, we calculated dN/dS, dI/dS, and dS from different sites as described in Methods. The dashed red line shows where dN/dS and dI/dS = 1, and therefore indicates neutrality.

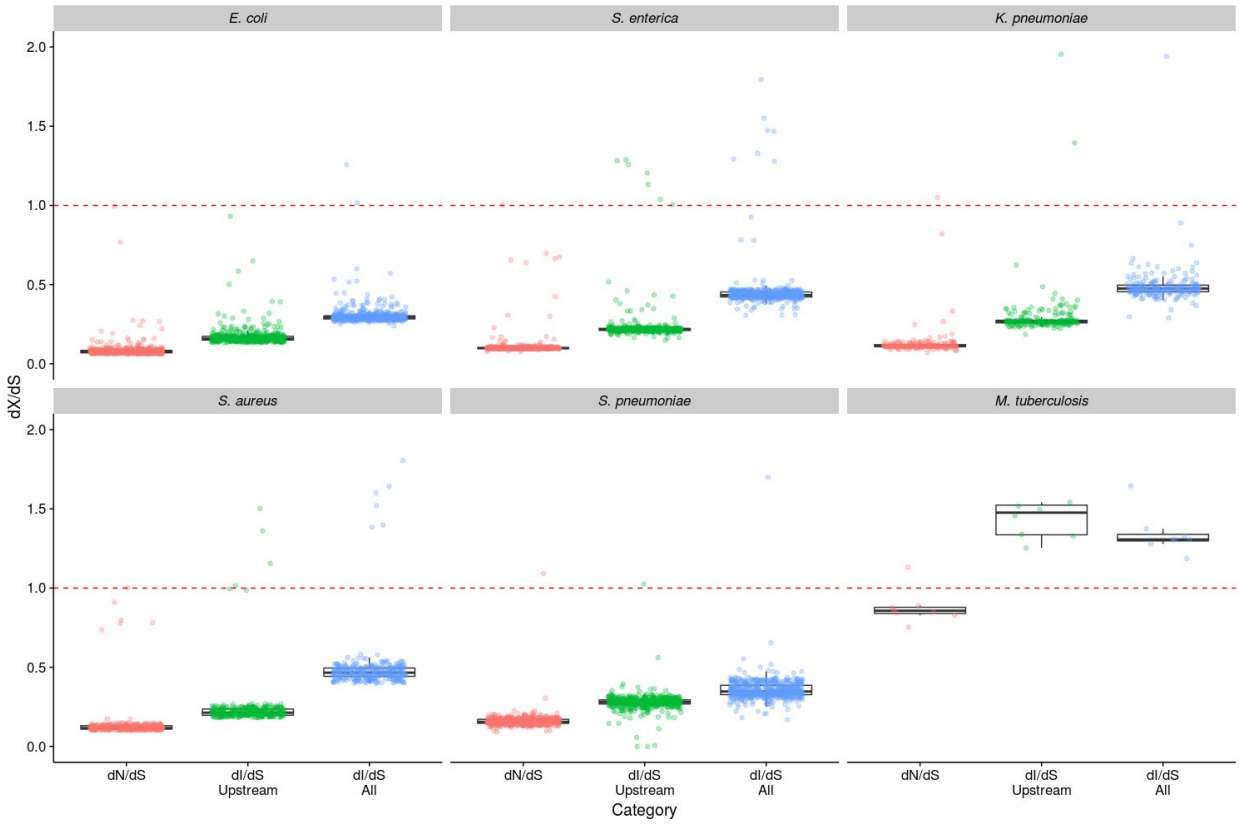

Figure S2.6: dI/dS analysis of IGRs upstream from gene starts. dI/dS was calculated from all intergenic sites, and intergenic sites 30 bp upstream from gene starts separately, by comparing isolates in a pairwise manner. The results were binned by dS (bin width = 0.0001) to control for oversampling of very closely related isolates (such as those belonging to the same CC). The genome-wide dN/dS values are included to enable comparisons to be made between non-synonymous sites and the intergenic sites. The dashed red line shows where dN/dS and dI/dS = 1, and therefore indicates neutrality.

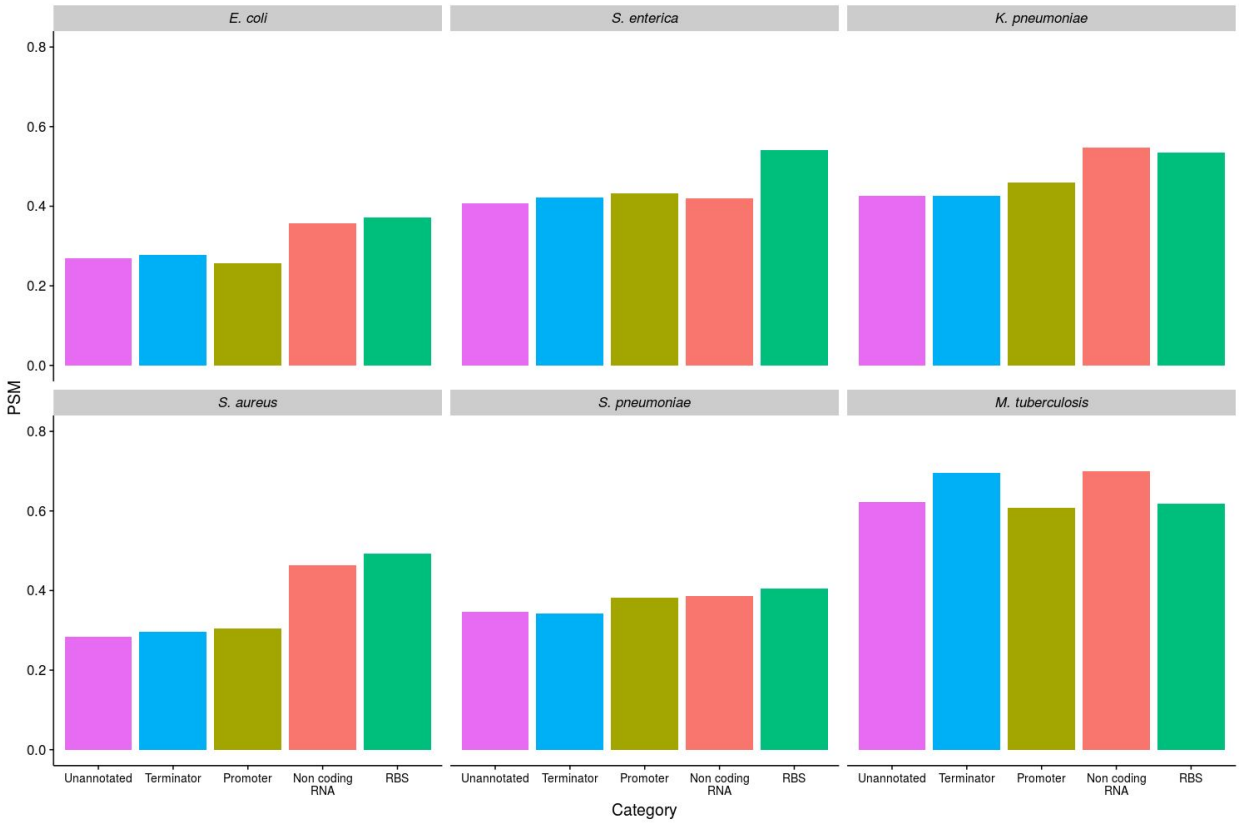

Figure S2.7: PSM analysis of selection on different regulatory elements within IGRs. PSM values were calculated by dividing the number of singleton SNPs by the total number of SNPs within that mutation category.

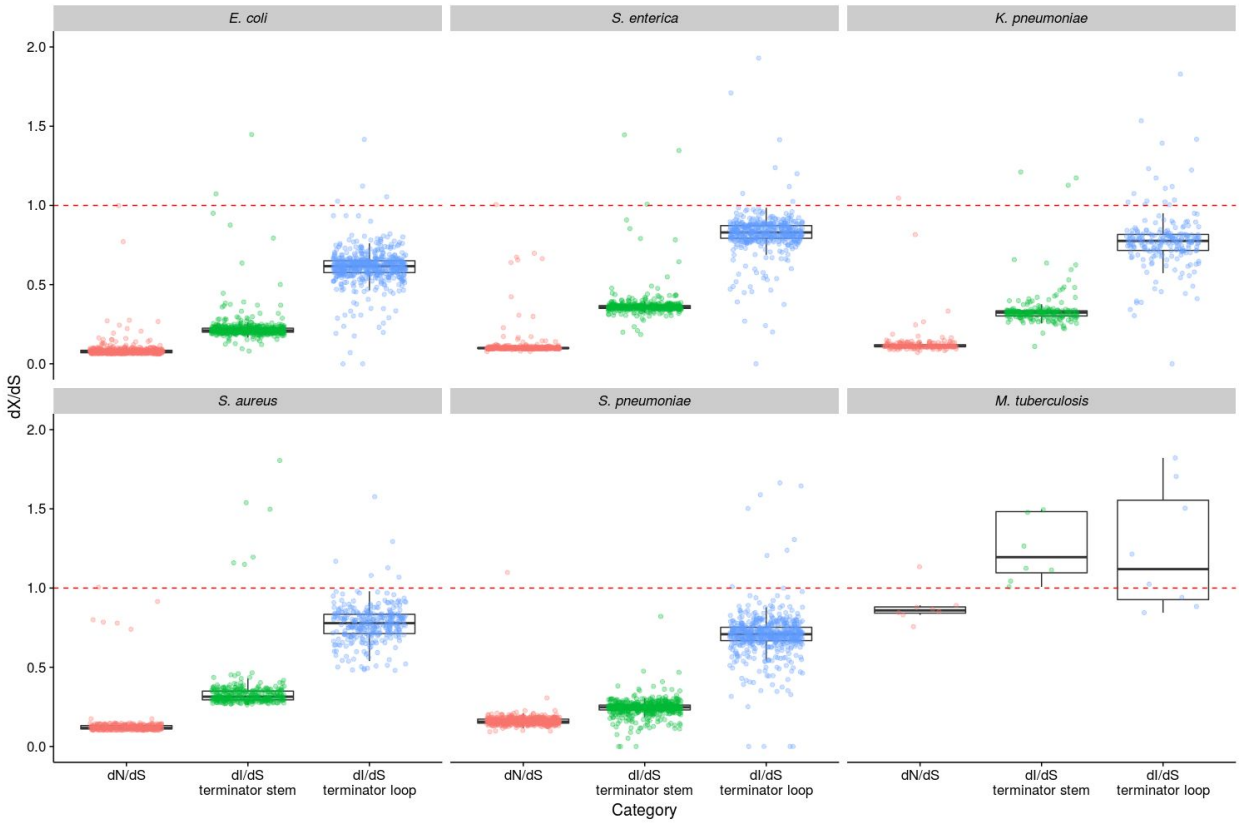

Figure S2.8: dI/dS analysis terminator stem and loop regions. dI/dS was calculated from terminator stem and loops separately, by comparing isolates in a pairwise manner. The results were binned by dS (bin width = 0.0001) to control for oversampling of very closely related isolates (such as those belonging to the same CC). The genome-wide dN/dS values are included to enable comparisons to be made between non-synonymous sites and the intergenic sites. The dashed red line shows where dN/dS and dI/dS = 1, and therefore indicates neutrality.

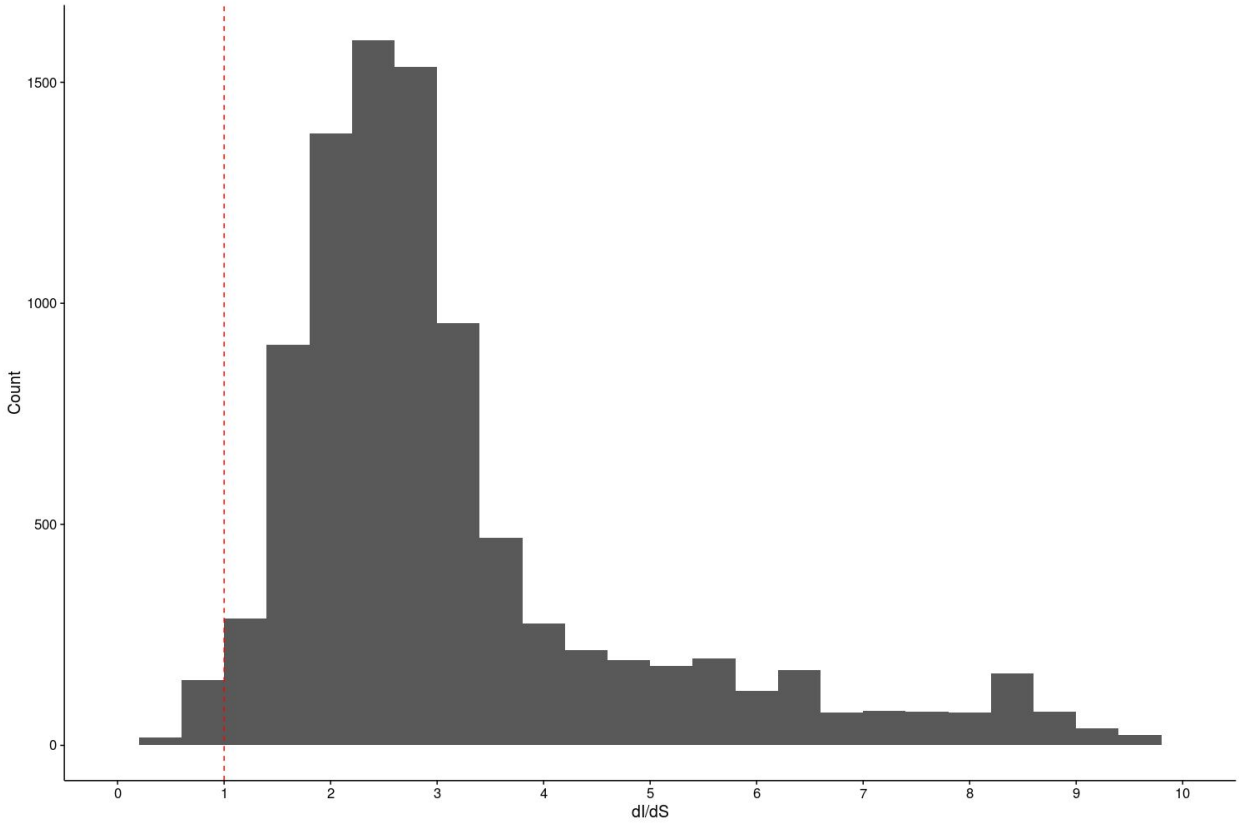

Figure S2.9: The distribution of promoter dI/dS values in *M. tuberculosis*. dI/dS was calculated from promoter sequences in a pairwise manner between *M. tuberculosis* isolates, and the histogram shows the distribution of these dI/dS values. 85% of the comparisons are > 1. The dashed red line shows where dI/dS = 1, and therefore indicates neutrality.

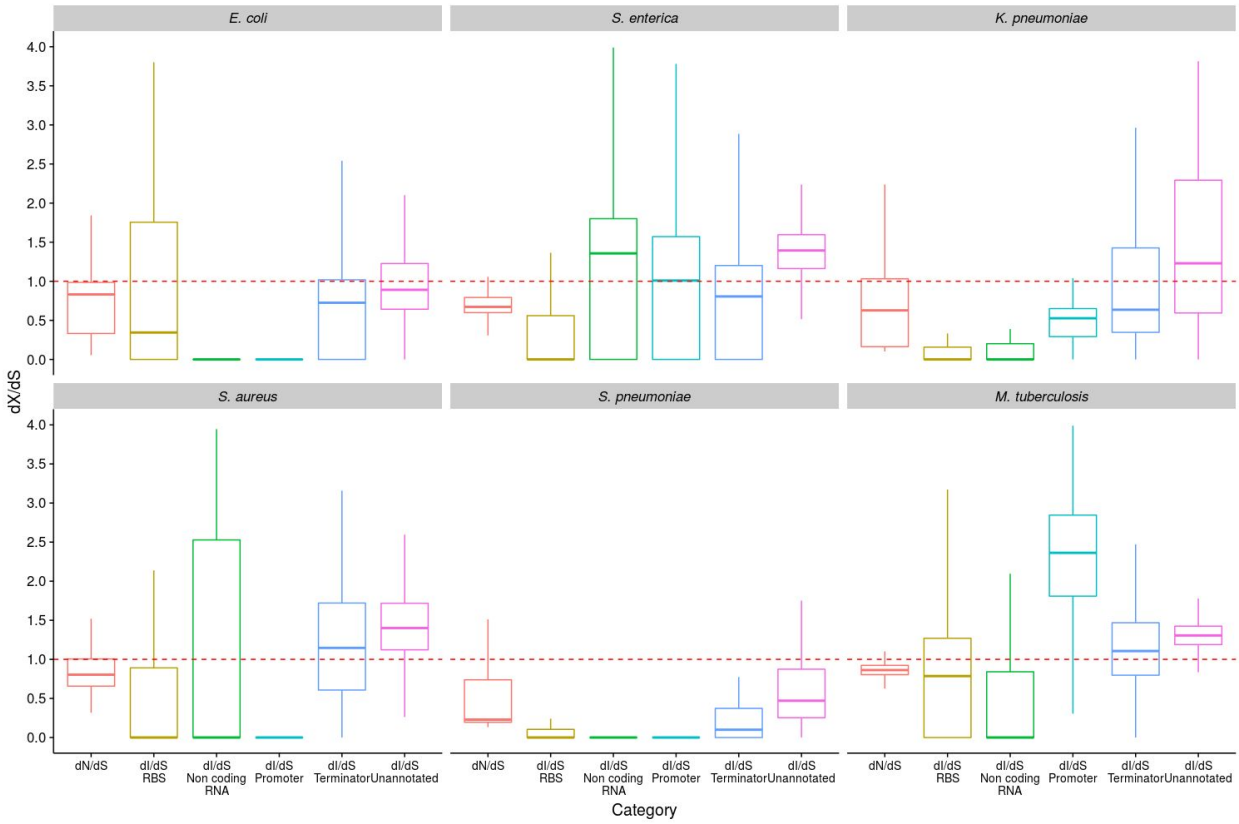

Figure S2.10: dI/dS analysis of selection on different regulatory elements in within-CC comparisons. dI/dS was calculated between isolates in a pairwise manner, and comparisons with dS < 0.003 were included to represent within-CC comparisons. The genome-wide dN/dS values are included to enable comparisons to be made between non-synonymous sites and the different regulatory intergenic sites. The dashed red line shows where dN/dS and dI/dS = 1, and therefore indicates neutrality.

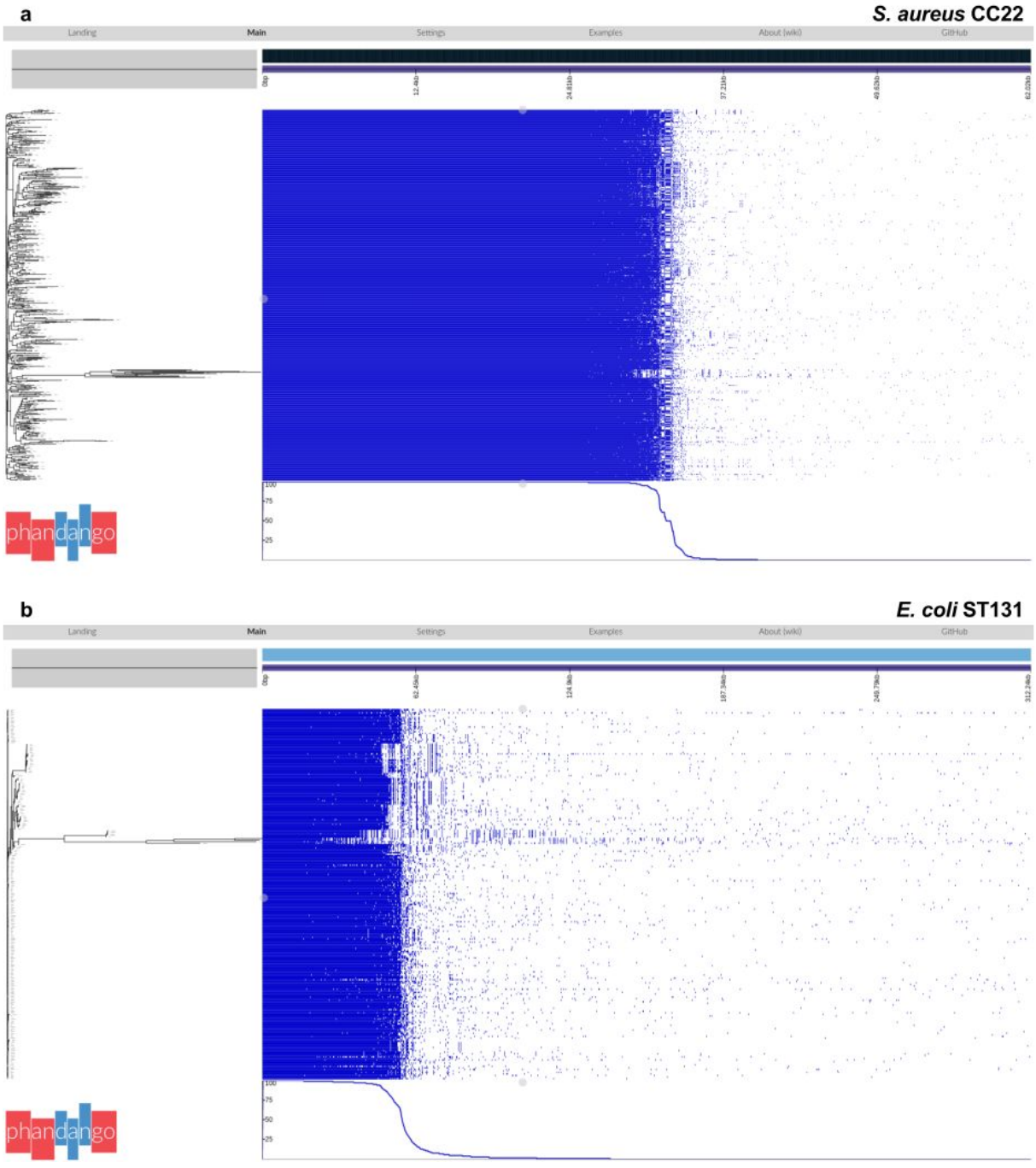

Figure S3.1: The IGR pan-genome ('panIGRome') as visualised using Phandango. A neighbour-joining phylogenetic tree was imported into Phandango alongside the IGR_presence_absence.csv file. Each row corresponds to an isolate, and each column corresponds to an IGR, with the IGRs ordered from the left in order of decreasing frequency within the sample. The line graph at the bottom shows the frequency of the IGRs within the sample. a) *S. aureus* ST22 b) *E. coli* ST131.

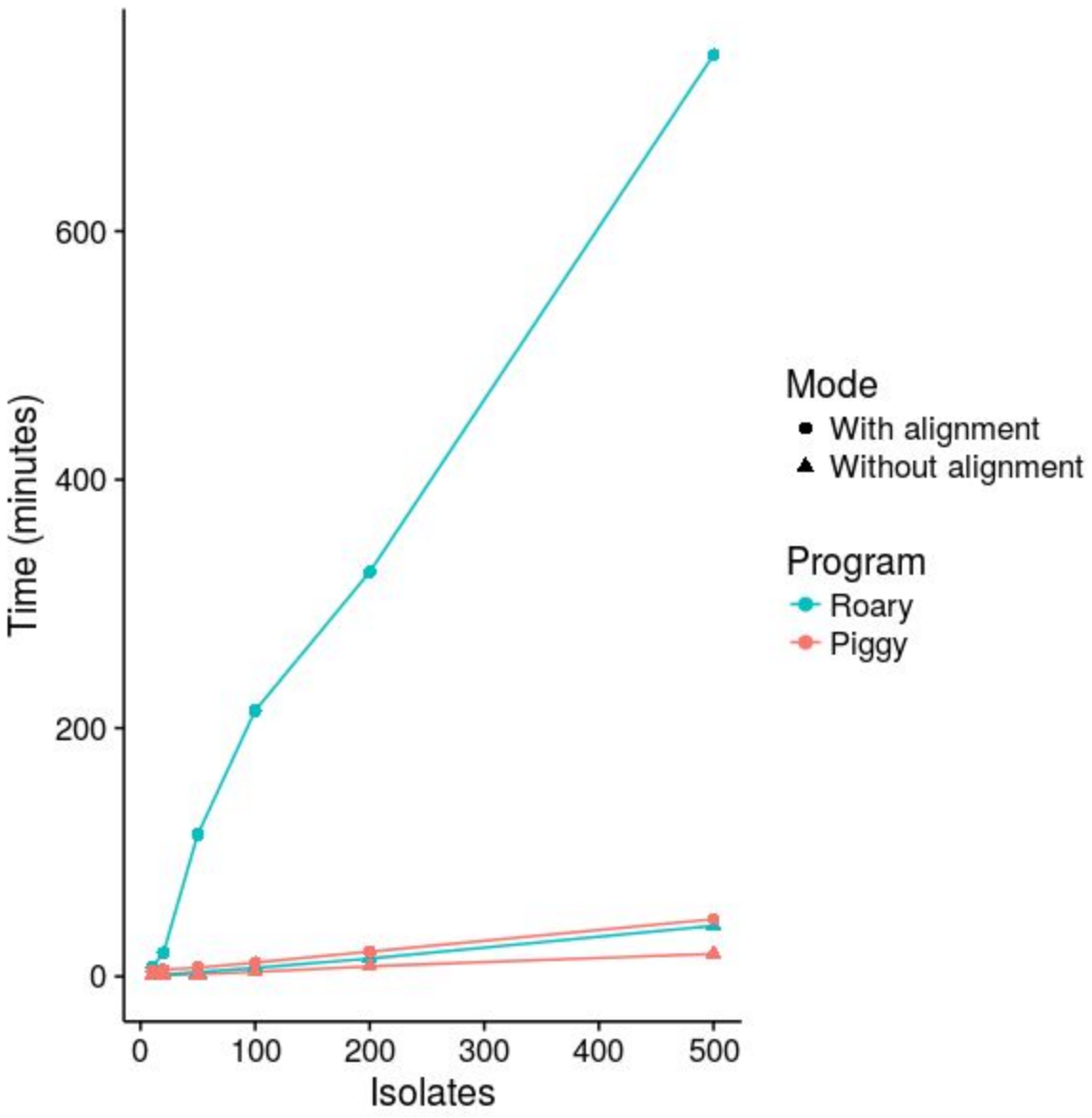

Figure S3.2: Comparison between Piggy and Roary. Roary (blue) and Piggy (red) were both run on increasing numbers of *S. aureus* ST22 isolates on a CLIMB virtual machine with 10 vcpus. The programs were both run with (circles) and without (triangles) alignment options.

Table S2.1: The isolates used in chapter 2.

Escherichia coli	MG1655	NC_004431	NC_002695	NC_012892	NC_007779	NC_011415
NC_013654	NC_013353	NC_013364	NC_007946	NC_008253	NC_008563	
NC_009801	NC_009800	NC_012967	NC_010468	NC_010473	NC_010498	
NC_011353	NC_013008	NC_012759	NC_012971	NC_017625	NC_012947	
NC_013941	NC_017628	NC_011748	NC_011741	CU928161	NC_011750	
NC_011601	NC_017626	NC_013361	NC_016902	NC_017631	NC_017632	
NC_017634	NC_017635	NC_017633	NC_017641	NC_017644	NZ_CP006632	
NZ_AGTD01000001	NC_017646	NC_017651	NC_017652	NC_017656	NC_017663	
NC_017660	NC_017664	NC_017906	NZ_AKBV01000001	NC_017638		
NZ_AKVX01000001	NC_011993	NZ_HG941718	NC_018650	NC_018658	NC_018661	
NC_020163	NC_020518	NC_022364	CP006698	NC_022370	NC_022648	
NZ_HG738867	NZ_CP006027	NZ_CP006262	NZ_CP007265	NZ_CP007390	NZ_CP007391	
NZ_CP007392	NZ_CP007393	NZ_CP007394	NZ_CP007133	NZ_CP007136	NZ_CP007799	
CP005998	NZ_CP008801	NZ_CP008805	NZ_CP008957	NZ_CP009072	NZ_CP009273	
NZ_CP009859	NZ_CP009644	NZ_CP009789	NZ_CP007149	NZ_CP009104	NZ_CP009106	
NZ_CP009685	NZ_CP010304	NZ_CP005930	NZ_CP010371	NZ_CP010315	NZ_CP007592	
NZ_CP009166	NZ_CP010585	NZ_CP010344	NZ_CP010816	NZ_CP010876	NZ_LM995446	
NZ_LM993812	NZ_HF572917	NZ_CP011134	NZ_CP011018	NZ_CP010438	NZ_CP010439	
NZ_CP010440	NZ_CP010441	NZ_CP010442	NZ_CP010443	NZ_CP010444	NZ_CP010445	
NZ_LN832404	NZ_CP011331	NZ_CP007594	NZ_CP011320	NZ_CP011321	CP011322	
CP011323	NZ_CP011324	NZ_CP011416	NZ_CP011342	NZ_CP011343	NZ_CP006636	
NZ_CP011938	NZ_CP011495	NZ_CP007442	NZ_CP012125	NZ_CP012126	NZ_CP012127	
NZ_CP011113	NZ_CP012635	NZ_CP012625	NZ_CP012633	NZ_CP012631	NZ_CP012802	
NZ_CP012868	NZ_CP012869	NZ_CP012870	NZ_CP013029	NZ_CP013025	NZ_CP013112	
NZ_CP013253	NZ_CP013658	NZ_CP008697	NZ_CP013831	NZ_CP013835	CP013837	
NZ_CP007491	NZ_CP014197	NZ_CP014225	NZ_CP014314	NZ_CP014268	NZ_CP014269	
NZ_CP014270	NZ_CP014272					

Salmonella enterica	Typhimurium_D23580	SRR1635091	SRR1635101	SRR1638968	
SRR1638975	SRR1638985	SRR1638992	SRR1638995	SRR1639000	SRR1639001
SRR1639006	SRR1639007	SRR1639009	SRR1643061	SRR1643065	SRR1643067
SRR1643076	SRR1643078	SRR1643079	SRR1643080	SRR1643082	SRR1643086
SRR1643090	SRR1643092	SRR1643094	SRR1643095	SRR1643101	SRR1643103
SRR1643104	SRR1643109	SRR1643112	SRR1643121	SRR1643123	SRR1643124
SRR1643129	SRR1643130	SRR1643131	SRR1643132	SRR1643133	SRR1643139
SRR1643141	SRR1643142	SRR1643144	SRR1643145	SRR1643148	SRR1643149
SRR1643151	SRR1643154	SRR1643155	SRR1643156	SRR1643157	SRR1643158

SRR1643159	SRR1643160	SRR1643161	SRR1644065	SRR1644066	SRR1644067
SRR1645070	SRR1645072	SRR1645074	SRR1645166	SRR1645181	SRR1645189
SRR1645193	SRR1645197	SRR1645198	SRR1645205	SRR1645211	SRR1645267
SRR1645279	SRR1645281	SRR1645343	SRR1645358	SRR1645363	SRR1645432
SRR1645441	SRR1645443	SRR1645457	SRR1645459	SRR1645462	SRR1645465
SRR1645535	SRR1645549	SRR1645564	SRR1645574	SRR1645586	SRR1645599
SRR1645601	SRR1645613	SRR1645617	SRR1645718	SRR1645736	SRR1645745
SRR1645758	SRR1645762	SRR1645772	SRR1645786	SRR1645795	SRR1645806
SRR1645829	SRR1645831	SRR1645849	SRR1645890	SRR1645891	SRR1645894
SRR1645915	SRR1645916	SRR1645924	SRR1645927	SRR1645929	SRR1645945
SRR1645955	SRR1645964	SRR1645965	SRR1645966	SRR1645970	SRR1645978
SRR1645983	SRR1645991	SRR1646024	SRR1646051	SRR1646052	SRR1646071
SRR1646098	SRR1646109	SRR1646113	SRR1646139	SRR1646160	SRR1646163
SRR1646166	SRR1646176	SRR1646197	SRR1646203	SRR1646226	SRR1646228
SRR1646245	SRR1646246	SRR1646255	SRR1646257	SRR1646261	SRR1646262
SRR1646271	SRR1646275	SRR1646276	SRR1646279	SRR1646283	SRR1646284
SRR1646287	SRR1646290	SRR1646349	SRR1646354	SRR1646361	SRR1646362
SRR1646366	SRR1646371	SRR1646372	SRR1646376	SRR1646379	SRR1646383
SRR1646389	SRR1646407	SRR1646408	SRR1646409	SRR1646410	SRR1952820
SRR1952885	SRR1957718	SRR1957733	SRR1957734	SRR1957742	SRR1957744
SRR1957768	SRR1957769	SRR1957772	SRR1957800	SRR1957805	SRR1957815
SRR1957826	SRR1957843	SRR1957848	SRR1957859	SRR1957867	SRR1957876
SRR1957887	SRR1957889	SRR1957901	SRR1957913	SRR1957919	SRR1957936
SRR1957937	SRR1957944	SRR1957945	SRR1957947	SRR1957962	SRR1957964
SRR1957970	SRR1957977	SRR1957985	SRR1957991	SRR1957998	SRR1958016
SRR1958031	SRR1958053	SRR1958063	SRR1958089	SRR1958091	SRR1958096
SRR1958099	SRR1958104	SRR1958125	SRR1958127	SRR1958147	SRR1958149
SRR1958174	SRR1958177	SRR1958178	SRR1958184	SRR1958211	SRR1958224
SRR1958245	SRR1958255	SRR1958263	SRR1958276	SRR1958277	SRR1958282
SRR1958294	SRR1958299	SRR1958300	SRR1958305	SRR1958311	SRR1958319
SRR1958323	SRR1958325	SRR1958326	SRR1958337	SRR1958357	SRR1958358
SRR1958365	SRR1958368	SRR1958376	SRR1958378	SRR1958380	SRR1958387
SRR1958401	SRR1958406	SRR1958425	SRR1958445	SRR1958446	SRR1958455
SRR1958471	SRR1958475	SRR1958480	SRR1958482	SRR1958506	SRR1958509
SRR1958531	SRR1958534	SRR1958543	SRR1958554	SRR1958563	SRR1958566
SRR1958567	SRR1958568	SRR1958569	SRR1958573	SRR1958585	SRR1958586
SRR1958593	SRR1958599	SRR1958602	SRR1958603	SRR1958604	SRR1958611
SRR1958614	SRR1958615	SRR1958622	SRR1958625	SRR1958627	SRR1958632
SRR1958637	SRR1958640	SRR1958643	SRR1958652	SRR1958655	SRR1958658

SRR1958663 SRR1958666 SRR1958667 SRR1958681 SRR1958686 SRR1959219
SRR1959220 SRR1959221 SRR1959225 SRR1959231 SRR1959235 SRR1959236
SRR1959240 SRR1959241 SRR1959242 SRR1959245 SRR1959264 SRR1959266
SRR1959268 SRR1959270 SRR1959277 SRR1959282 SRR1959283 SRR1959290
SRR1959291 SRR1959293 SRR1959304 SRR1959307 SRR1959309 SRR1959310
SRR1959313 SRR1959317 SRR1959319 SRR1959364 SRR1959365 SRR1959367
SRR1959368 SRR1959394 SRR1959401 SRR1959402 SRR1959410 SRR1959413
SRR1959414 SRR1959418 SRR1959430 SRR1959442 SRR1959445 SRR1959447
SRR1959449 SRR1959450 SRR1959456 SRR1959466 SRR1959474 SRR1959483
SRR1959491 SRR1959494 SRR1959718 SRR1959722 SRR1959725 SRR1960028
SRR1960030 SRR1960039 SRR1960046 SRR1960047 SRR1960055 SRR1960059
SRR1960070 SRR1960079 SRR1960087 SRR1960121 SRR1960130 SRR1960133
SRR1960137 SRR1960142 SRR1960154 SRR1960161 SRR1960183 SRR1960195
SRR1960196 SRR1960201

Klebsiella pneumoniae NTUH_K2044 ERR024819 ERR024821 ERR024822 ERR024823
ERR024824 ERR024831 ERR024832 ERR024835 ERR024836 ERR024837
ERR024839 ERR024840 ERR024843 ERR024845 ERR024848 ERR024851
ERR024854 ERR025098 ERR025100 ERR025103 ERR025107 ERR025109
ERR025111 ERR025113 ERR025115 ERR025116 ERR025117 ERR025118
ERR025119 ERR025121 ERR025125 ERR025128 ERR025135 ERR025137
ERR025139 ERR025140 ERR025141 ERR025142 ERR025143 ERR025144
ERR025146 ERR025147 ERR025150 ERR025151 ERR025152 ERR025154
ERR025156 ERR025159 ERR025160 ERR025462 ERR025464 ERR025465
ERR025468 ERR025469 ERR025470 ERR025471 ERR025472 ERR025473
ERR025475 ERR025477 ERR025478 ERR025479 ERR025482 ERR025483
ERR025484 ERR025485 ERR025486 ERR025489 ERR025492 ERR025494
ERR025495 ERR025496 ERR025497 ERR025498 ERR025499 ERR025501
ERR025502 ERR025503 ERR025504 ERR025505 ERR025506 ERR025510
ERR025511 ERR025512 ERR025515 ERR025516 ERR025517 ERR025518
ERR025519 ERR025520 ERR025521 ERR025522 ERR025523 ERR025524
ERR025525 ERR025527 ERR025529 ERR025531 ERR025532 ERR025534
ERR025535 ERR025536 ERR025538 ERR025540 ERR025541 ERR025542
ERR025543 ERR025544 ERR025545 ERR025546 ERR025547 ERR025548
ERR025550 ERR025553 ERR025554 ERR025555 ERR025557 ERR025558
ERR025561 ERR025562 ERR025563 ERR025564 ERR025566 ERR025570
ERR025571 ERR025572 ERR025574 ERR025575 ERR025576 ERR025581
ERR025583 ERR025585 ERR025589 ERR025590 ERR025593 ERR025594
ERR025595 ERR025596 ERR025597 ERR025600 ERR025601 ERR025602

ERR025603	ERR025605	ERR025606	ERR025607	ERR025609	ERR025610
ERR025614	ERR025616	ERR025618	ERR025619	ERR025620	ERR025621
ERR025622	ERR025624	ERR025627	ERR025628	ERR025633	ERR025634
ERR025635	ERR025636	ERR025637	ERR025638	ERR025641	ERR025642
ERR025645	ERR025646	ERR025647	ERR025648	ERR025649	ERR025652
ERR025654	ERR025655	ERR025657	ERR025658	ERR025659	ERR025660
ERR025661	ERR025663	ERR025664	ERR025667	ERR025668	ERR025670
ERR025672	ERR025673	ERR025674	ERR025675	ERR025677	ERR025678
ERR025679	ERR025680	ERR025979	ERR025980	ERR025981	ERR025983
ERR025984	ERR025985	ERR025988	ERR025989	ERR025992	ERR025993
ERR025994	ERR025996	ERR025998	ERR026000	ERR026001	

Staphylococcus aureus	HO_5096_0412	ERR084677	ERR084668	ERR084649	ERR084554
ERR084555	ERR084492	ERR114901	ERR109598	ERR114881	ERR109499
ERR109494	ERR109523	ERR114857	ERR109681	ERR114861	ERR109589
ERR109575	ERR114910	ERR109483	ERR172072	ERR172073	ERR172075
ERR109685	ERR109532	ERR223120	ERR109555	ERR109691	ERR172035
ERR109519	ERR114913	ERR172065	ERR109654	ERR223176	ERR223173
ERR223171	ERR114871	ERR172025	ERR223117	ERR109625	ERR109536
ERR114903	ERR172055	ERR109661	ERR109663	ERR109631	ERR109634
ERR109628	ERR109581	ERR109558	ERR114897	ERR114894	ERR107843
ERR118523	ERR223143	ERR223140	ERR223141	ERR118411	ERR118414
ERR158654	ERR158699	ERR107942	ERR107902	ERR129301	ERR111129
ERR107844	ERR107938	ERR118336	ERR118480	ERR107950	ERR158798
ERR134385	ERR124474	ERR129326	ERR107834	ERR118579	ERR118580
ERR118353	ERR118330	ERR158729	ERR107913	ERR107792	ERR158776
ERR124530	ERR118468	ERR129333	ERR134397	ERR134395	ERR134399
ERR134352	ERR111090	ERR124456	ERR129321	ERR118385	ERR158631
ERR118462	ERR118605	ERR107974	ERR134370	ERR134368	ERR124492
ERR158727	ERR129264	ERR107799	ERR158691	ERR111058	ERR118514
ERR134335	ERR134345	ERR118377	ERR158753	ERR158756	ERR107854
ERR134402	ERR118446	ERR118454	ERR107946	ERR156430	ERR156433
ERR156496	ERR156497	ERR159006	ERR159008	ERR223164	ERR156493
ERR156510	ERR158985	ERR156498	ERR156509	ERR177165	ERR159050
ERR156518					

Streptococcus pneumoniae	ATCC_700669	ERR039573	ERR039578	ERR039560	
ERR039563	ERR039567	ERR039605	ERR039585	ERR039592	ERR039620
ERR039621	ERR039622	ERR039629	ERR039631	ERR047914	ERR047916

ERR047895	ERR047922	ERR047925	ERR047954	ERR047956	ERR047947
ERR047980	ERR047981	ERR047983	ERR047989	ERR047969	ERR048024
ERR048031	ERR048014	ERR048094	ERR048121	ERR048128	ERR048131
ERR048114	ERR048118	ERR048151	ERR048156	ERR048135	ERR048137
ERR048141	ERR048176	ERR048196	ERR048199	ERR048205	ERR048187
ERR050018	ERR050044	ERR051421	ERR051432	ERR051415	ERR051450
ERR051442	ERR049091	ERR049092	ERR049082	ERR049115	ERR049949
ERR049955	ERR049937	ERR052579	ERR050071	ERR050076	ERR050079
ERR052608	ERR052635	ERR052627	ERR273512	ERR051471	ERR051461
ERR051462	ERR051483	ERR051484	ERR051488	ERR051512	ERR051513
ERR051540	ERR051549	ERR051534	ERR051535	ERR051559	ERR273539
ERR051625	ERR051610	ERR051637	ERR051643	ERR051648	ERR051628
ERR273494	ERR273496	ERR273497	ERR273500	ERR273485	ERR054228
ERR054265	ERR054267	ERR054250	ERR051591	ERR051593	ERR051579
ERR051585	ERR054303	ERR054304	ERR054307	ERR054330	ERR054333
ERR054359	ERR054362	ERR054350	ERR054380	ERR054382	ERR054433
ERR054419	ERR054450	ERR054455	ERR054442	ERR054473	ERR054468
ERR054505	ERR054509	ERR054516	ERR054549	ERR054533	ERR054556
ERR054574	ERR054575	ERR054597	ERR054598	ERR054614	ERR054617
ERR054626	ERR054648	ERR056716	ERR056708	ERR056711	ERR056687
ERR056772	ERR056774	ERR056775	ERR056757	ERR056759	ERR056760
ERR056762	ERR056763	ERR056793	ERR056779	ERR056802	ERR056814
ERR056815	ERR056803	ERR056808	ERR056830	ERR056833	ERR056865
ERR056851	ERR056852	ERR056855	ERR056876	ERR056879	ERR056880
ERR057772	ERR057778	ERR057782	ERR057790	ERR057821	ERR057826
ERR057809	ERR057814	ERR059994	ERR060004	ERR060025	ERR060026
ERR060007	ERR060043	ERR060032	ERR063808	ERR063821	ERR063846
ERR063841	ERR063868	ERR063877	ERR063866	ERR063896	ERR063890
ERR063906	ERR063924	ERR063925	ERR063952	ERR063982	ERR064033
ERR064034	ERR064057	ERR066191	ERR066192	ERR066197	ERR066183
ERR064110	ERR064096	ERR064133	ERR064136	ERR064169	ERR064202
ERR064204	ERR064205	ERR066216	ERR066228	ERR066236	ERR066249
ERR066253	ERR066270	ERR066271	ERR066290	ERR066279	ERR066299
ERR066325	ERR066331	ERR066361	ERR067881	ERR067865	ERR067905
ERR067907	ERR067954	ERR067958	ERR069622	ERR069647	ERR069653
ERR069639	ERR069674	ERR069664	ERR084168	ERR084190	ERR084185
ERR084217	ERR084248	ERR084229	ERR084252	ERR084257	ERR072192
ERR072195	ERR072182	ERR072213	ERR072219	ERR072220	ERR072244
ERR084296	ERR084314	ERR057901	ERR057904	ERR057906	ERR057909

ERR057941	ERR057963	ERR057966	ERR057977	ERR057978	ERR067777
ERR067789	ERR067808	ERR067811	ERR067795	ERR067825	ERR067826
ERR067820	ERR063859				

Mycobacterium tuberculosis	H37Rv	ERR017792	ERR017798	ERR019568	ERR019569
ERR019573	ERR019852	ERR019871	ERR019875	ERR026636	ERR027462
ERR027466	ERR047884	ERR047890	ERR067606	ERR067607	ERR067628
ERR067632	ERR067637	ERR067656	ERR067659	ERR067671	ERR067674
ERR067677	ERR067682	ERR067686	ERR067703	ERR108424	ERR108439
ERR108442	ERR108453	ERR108458	ERR108463	ERR108480	ERR108482
ERR117449	ERR117465	ERR117467	ERR133801	ERR133812	ERR133847
ERR133851	ERR133858	ERR133870	ERR133871	ERR133892	ERR133897
ERR133901	ERR133902	ERR133905	ERR133913	ERR133918	ERR133919
ERR133924	ERR133940	ERR133947	ERR133951	ERR133954	ERR133963
ERR133981	ERR137208	ERR137213	ERR137217	ERR137219	ERR137224
ERR137225	ERR137229	ERR137247	ERR137251	ERR137266	ERR137280
ERR137281	ERR144546	ERR144551	ERR144556	ERR144567	ERR144572
ERR144573	ERR144576	ERR144579	ERR144600	ERR144602	ERR144615
ERR144625	ERR158570	ERR158580	ERR158585	ERR158586	ERR158592
ERR158600	ERR158603	ERR158610	ERR158612	ERR227978	ERR228018
ERR228020	ERR228021	ERR228025	ERR228026	ERR228045	ERR228057
ERR228062	ERR228066	ERR229937	ERR229940	ERR229945	ERR229952
ERR229958	ERR229972	ERR229973	ERR229974	ERR229979	ERR229983
ERR229992	ERR230000	ERR230006	ERR234561	ERR234575	ERR234587
ERR234597	ERR234614	ERR234621	ERR234627	ERR234638	ERR234642
ERR234672	ERR234675	ERR234676	ERR234678	ERR234681	ERR234683
ERR234690	ERR234695	ERR403216	ERR403246	ERR403247	ERR403252
ERR403254	ERR403257	ERR403273	ERR403274	ERR403289	ERR403312
ERR403314					

Table S2.2: Information about genes with promoter SNPs in *M. tuberculosis*.

Genomic_position	Number_of_isolates_with_SNP	Gene_Prokka_id	Gene	Product
27463	4	M_tuberculosis_00030		putative_HTH-type_transcriptional_regulator/MT0026
27487	5	M_tuberculosis_00030		putative_HTH-type_transcriptional_regulator/MT0026
66827	1	M_tuberculosis_00070	yvdP	putative_FAD-linked_oxidoreductase_YvdP
136132	1	M_tuberculosis_00125	rmd	GDP-6-deoxy-D-mannose_reductase
163320	1	M_tuberculosis_00149		Bacterial_regulatory_proteins%2C_tetR_family
277865	1	M_tuberculosis_00252		DNA-binding_transcriptional_regulator_EnvR
394059	1	M_tuberculosis_00351	ptlI_2	Pentalenene_oxygenase
684376	3	M_tuberculosis_00624	mce2R_2	HTH-type_transcriptional_regulator_Mce2R
696916	1	M_tuberculosis_00637		hypothetical_protein
716376	1	M_tuberculosis_00665		Antitoxin_VapB30
850239	5	M_tuberculosis_00804		putative_PPE_family_protein_PPE42
886661	13	M_tuberculosis_00846		Putative_monooxygenase
919563	1	M_tuberculosis_00881		hypothetical_protein
919564	1	M_tuberculosis_00881		hypothetical_protein
935511	1	M_tuberculosis_00898	ubiE_2	Demethylmenaquinone_methyltransferase
944296	4	M_tuberculosis_00905		hypothetical_protein
944316	1	M_tuberculosis_00905		hypothetical_protein
986515	1	M_tuberculosis_00947		hypothetical_protein
993795	2	M_tuberculosis_00952	hapE_1	4-hydroxyacetophenone_monooxygenase
1021717	1	M_tuberculosis_00978		PE_family_protein
1163808	3	M_tuberculosis_01108		PE_family_protein
1172378	2	M_tuberculosis_01118	slyA_1	Transcriptional_regulator_SlyA
1241572	3	M_tuberculosis_01191		Antibiotic_biosynthesis_monooxygenase
1254543	1	M_tuberculosis_01205	prpD	2-methylcitrate_dehydratase
1348638	1	M_tuberculosis_01286		hypothetical_protein
1525196	1	M_tuberculosis_01446		hypothetical_protein
1708792	3	M_tuberculosis_01615	epsE_1	Putative_glycosyltransferase_EpsE
1728603	50	M_tuberculosis_01626	pks2_3	Phthioceranic/hydroxyphthioceranic_acid_synthase
1728622	16	M_tuberculosis_01626	pks2_3	Phthioceranic/hydroxyphthioceranic_acid_synthase
1733563	1	M_tuberculosis_01631		hypothetical_protein
1774847	1	M_tuberculosis_01667	ripA_2	Peptidoglycan_endopeptidase_RipA_precursor
1855659	1	M_tuberculosis_01749		PE_family_protein
1906307	1	M_tuberculosis_01787		DivIVA_protein
2056377	1	M_tuberculosis_01930		Fatty_acid_hydroxylase_superfamily_protein
2135870	14	M_tuberculosis_02004		hypothetical_protein
2170943	1	M_tuberculosis_02040		putative_PPE_family_protein_PPE42

21770491	M_tuberculosis_02046		hypothetical_protein
217707312	M_tuberculosis_02046		hypothetical_protein
21873062	M_tuberculosis_02057	fcbB2	4-chlorobenzoyl_coenzyme_A_dehalogenase-2
222536569	M_tuberculosis_02108		hypothetical_protein
22299611	M_tuberculosis_02114		putative_HTH-type_transcriptional_regulator/MT2039
22380101	M_tuberculosis_02126	cmtR	HTH-type_transcriptional_regulator_CmtR
22380331	M_tuberculosis_02126	cmtR	HTH-type_transcriptional_regulator_CmtR
226505913	M_tuberculosis_02151		hypothetical_protein
22651981	M_tuberculosis_02151		hypothetical_protein
22718321	M_tuberculosis_02158		putative_cation_efflux_system_protein/MT2084
22812781	M_tuberculosis_02166		hypothetical_protein
22812791	M_tuberculosis_02166		hypothetical_protein
23269091	M_tuberculosis_02204	sigC	ECF_RNA_polymerase_sigma_factor_SigC
26107021 6	M_tuberculosis_02492		hypothetical_protein
26843461	M_tuberculosis_02549		hypothetical_protein
29279392	M_tuberculosis_02764		hypothetical_protein
29486311	M_tuberculosis_02786		Helix-turn-helix_domain_protein
3086742 10	M_tuberculosis_02944	ald	Alanine_dehydrogenase
30867471	M_tuberculosis_02944	ald	Alanine_dehydrogenase
3086788 18	M_tuberculosis_02944	ald	Alanine_dehydrogenase
33631851	M_tuberculosis_03178	ilvB1	Acetolactate_synthase_large_subunit_IlvB1
33810821	M_tuberculosis_03199		PE_family_protein
34194679	M_tuberculosis_03237		Putative_cytochrome_P450_120
36402891	M_tuberculosis_03457	whiB2	Transcriptional_regulator_WhiB2
36404081	M_tuberculosis_03457	whiB2	Transcriptional_regulator_WhiB2
40564531	M_tuberculosis_03827	espA_1	ESX-1_secretion-associated_protein_EspA
40570363	M_tuberculosis_03828		Soluble_epoxide_hydrolase
41216742	M_tuberculosis_03893	whiB4	Transcriptional_regulator_WhiB4
41216757	M_tuberculosis_03893	whiB4	Transcriptional_regulator_WhiB4
41494882	M_tuberculosis_03919		hypothetical_protein
41953821	M_tuberculosis_03962	ctpJ	putative_cation-transporting_P-type_ATPase_J
41953903	M_tuberculosis_03962	ctpJ	putative_cation-transporting_P-type_ATPase_J
42054581	M_tuberculosis_03980	osmF	
Putative_osmoprotectant_uptake_system_substrate-binding_protein_OsmF_precursor			
432748015	M_tuberculosis_04086	ethA_4	FAD-containing_monooxygenase_EthA
433659712	M_tuberculosis_04092	gltB_2	Glutamate_synthase_[NADPH]_large_chain

Supplementary form SF1.

This declaration concerns the article entitled:									
Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species									
Publication status (tick one)									
draft manuscript		**Submitted**		**In review**		**Accepted**		**Published**	X
Publication details (reference)	Thorpe HA., Bayliss SC., Hurst LD., and Feil EJ. 2017. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. Genetics 206 (1): 363-76.								
Candidate's contribution to the paper (detailed, and also given as a percentage).	The candidate contributed to/ considerably contributed to/predominantly executed the… Formulation of ideas: Considerably contributed to 85% Design of methodology: Predominantly executed the 90% Experimental work: Predominantly executed the 95% Presentation of data in journal format: Predominantly executed the 90%								
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.								
Signed						**Date**	29/09/2017		

Supplementary form SF2.

This declaration concerns the article entitled:									
Piggy: A rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria									
Publication status (tick one)									
draft manuscript		**Submitted**		**In review**	X	**Accepted**		**Published**	
Publication details (reference)	Thorpe HA., Bayliss SC., Sheppard SK., and Feil EJ. Piggy: A rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. bioRxiv. doi:10.1101/179515								
Candidate's contribution to the paper (detailed, and also given as a percentage).	The candidate contributed to/ considerably contributed to/predominantly executed the… Formulation of ideas: Considerably contributed to 85% Design of methodology: Predominantly executed the 90% Experimental work: Predominantly executed the 95% Presentation of data in journal format: Predominantly executed the 90%								
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.								
Signed						**Date**	29/09/2017		

References

Aanensen, David M., Edward J. Feil, Matthew T. G. Holden, Janina Dordel, Corin A. Yeats, Artemij Fedosejev, Richard Goater, et al. 2016. “Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: A Population Snapshot of Invasive Staphylococcus Aureus in Europe.” *mBio* 7 (3). https://doi.org/10.1128/mBio.00444-16.

Acebo, P., A. J. Martin-Galiano, S. Navarro, A. Zaballos, and M. Amblar. 2012. “Identification of 88 Regulatory Small RNAs in the TIGR4 Strain of the Human Pathogen Streptococcus Pneumoniae.” *RNA*.

Allardet-Servent, A., S. Michaux-Charachon, E. Jumas-Bilak, L. Karayan, and M. Ramuz. 1993. “Presence of One Linear and One Circular Chromosome in the Agrobacterium Tumefaciens C58 Genome.” *Journal of Bacteriology* 175 (24):7869–74.

Andreani, Nadia Andrea, Elze Hesse, and Michiel Vos. 2017. “Prokaryote Genome Fluidity Is Dependent on Effective Population Size.” *The ISME Journal* 11 (7):1719–21.

Balbi, Kevin J., Eduardo P. C. Rocha, and Edward J. Feil. 2009. “The Temporal Dynamics of Slightly Deleterious Mutations in Escherichia Coli and Shigella Spp.” *Molecular Biology and Evolution* 26 (2):345–55.

Bennett, Gordon M., and Nancy A. Moran. 2013. “Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect.” *Genome Biology and Evolution* 5 (9):1675–88.

Botella, Laure, Julien Vaubourgeix, Jonathan Livny, and Dirk Schnappinger. 2017. “Depleting Mycobacterium Tuberculosis of the Transcription Termination Factor Rho Causes Pervasive Transcription and Rapid Death.” *Nature Communications* 8 (March):14731.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5):525–27.

Brochet, Mathieu, Christophe Rusniok, Elisabeth Couvé, Shaynoor Dramsi, Claire Poyart, Patrick Trieu-Cuot, Frank Kunst, and Philippe Glaser. 2008. “Shaping a Bacterial Genome by Large Chromosomal Replacements, the Evolutionary History of Streptococcus Agalactiae.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (41):15961–66.

Browning, Douglas F., and Stephen J. Busby. 2004. “The Regulation of Bacterial Transcription Initiation.” *Nature Reviews. Microbiology* 2 (1):57–65.

Browning, Douglas F., and Stephen J. W. Busby. 2016. “Local and Global Regulation of

Transcription Initiation in Bacteria." *Nature Reviews. Microbiology* 14 (10). Nature Research:638–50.

Brynildsrud, Ola, Jon Bohlin, Lonneke Scheffer, and Vegard Eldholm. 2016. "Rapid Scoring of Genes in Microbial Pan-Genome-Wide Association Studies with Scoary." *Genome Biology* 17 (1):238.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (December):421.

Canchaya, Carlos, Ghislain Fournous, Sandra Chibani-Chennoufi, Marie Lise Dillmann, and Harald Brüssow. 2003. "Phage as Agents of Lateral Gene Transfer." *Current Opinion in Microbiology* 6 (4):417–24.

Casali, Nicola, Vladyslav Nikolayevskyy, Yanina Balabanova, Simon R. Harris, Olga Ignatyeva, Irina Kontsevaya, Jukka Corander, et al. 2014. "Evolution and Transmission of Drug-Resistant Tuberculosis in a Russian Population." *Nature Genetics* 46 (3):279–86.

Casali, Nicola, Vladyslav Nikolayevskyy, Yanina Balabanova, Olga Ignatyeva, Irina Kontsevaya, Simon R. Harris, Stephen D. Bentley, et al. 2012. "Microevolution of Extensively Drug-Resistant Tuberculosis in Russia." *Genome Research* 22 (4):735–45.

Casillas, Sònia, and Antonio Barbadilla. 2017. "Molecular Population Genetics." *Genetics* 205 (3):1003–35.

Castillo, Andrea R., Sonia S. Arevalo, Andrew J. Woodruff, and Karen M. Ottemann. 2008. "Experimental Analysis of Helicobacter Pylori Transcriptional Terminators Suggests This Microbe Uses Both Intrinsic and Factor-Dependent Termination." *Molecular Microbiology* 67 (1):155–70.

Castillo-Ramírez, Santiago, Simon R. Harris, Matthew T. G. Holden, Miao He, Julian Parkhill, Stephen D. Bentley, and Edward J. Feil. 2011. "The Impact of Recombination on dN/dS within Recently Emerged Bacterial Clones." *PLoS Pathogens* 7 (7).

Chaguza, Chrispin, Cheryl P. Andam, Simon R. Harris, Jennifer E. Cornick, Marie Yang, Laura Bricio-Moreno, Arox W. Kamng'ona, et al. 2016. "Recombination in Streptococcus Pneumoniae Lineages Increase with Carriage Duration and Size of the Polysaccharide Capsule." *mBio* 7 (5). https://doi.org/10.1128/mBio.01053-16.

Chan, Jacqueline Z-M, Mark J. Pallen, Beryl Oppenheim, and Chrystala Constantinidou. 2012. "Genome Sequencing in Clinical Microbiology." *Nature Biotechnology* 30 (11):1068–71.

Charlesworth, Brian. 2009. "Fundamental Concepts in Genetics: Effective Population Size and

Patterns of Molecular Evolution and Variation." *Nature Reviews. Genetics* 10 (3):195–205.

Charneski, Catherine A., Frank Honti, Josephine M. Bryant, Laurence D. Hurst, and Edward J. Feil. 2011. "Atypical AT Skew in Firmicute Genomes Results from Selection and Not from Mutation." *PLoS Genetics* 7 (9).

Chauhan, Santosh, Anil Kumar, Amit Singhal, Jaya Sivaswami Tyagi, and H. Krishna Prasad. 2009. "CmtR, a Cadmium-Sensing ArsR-SmtB Repressor, Cooperatively Interacts with Multiple Operator Sites to Autorepress Its Transcription in Mycobacterium Tuberculosis." *The FEBS Journal* 276 (13):3428–39.

Chen, Xiaoshu, and Jianzhi Zhang. 2013. "No Gene-Specific Optimization of Mutation Rate in Escherichia Coli." *Molecular Biology and Evolution* 30 (7):1559–62.

Chewapreecha, Claire, Simon R. Harris, Nicholas J. Croucher, Claudia Turner, Pekka Marttinen, Lu Cheng, Alberto Pessia, et al. 2014. "Dense Genomic Sampling Identifies Highways of Pneumococcal Recombination." *Nature Genetics* 46 (3):305–9.

Ciampi, M. Sofia. 2006. "Rho-Dependent Terminators and Transcription Termination." *Microbiology* 152 (Pt 9):2515–28.

Collins, Caitlin, and Xavier Didelot. 2017. "A Phylogenetic Method To Perform Genome-Wide Association Studies In Microbes That Accounts For Population Structure And Recombination." *bioRxiv*. https://doi.org/10.1101/140798.

Connor, Thomas R., Nicholas J. Loman, Simon Thompson, Andy Smith, Joel Southgate, Radoslaw Poplawski, Matthew J. Bull, et al. 2016. "CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): An Online Resource for the Medical Microbiology Community." *Microbial Genomics* 2 (9). Microbiology Society. https://doi.org/10.1099/mgen.0.000086.

Cooper, Vaughn S., Samuel H. Vohr, Sarah C. Wrocklage, and Philip J. Hatcher. 2010. "Why Genes Evolve Faster on Secondary Chromosomes in Bacteria." *PLoS Computational Biology* 6 (4):e1000732.

Croucher, Nicholas J., Rafal Mostowy, Christopher Wymant, Paul Turner, Stephen D. Bentley, and Christophe Fraser. 2016. "Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict." *PLoS Biology* 14 (3). Public Library of Science:e1002394.

Cui, Yujun, Xianwei Yang, Xavier Didelot, Chenyi Guo, Dongfang Li, Yanfeng Yan, Yiquan Zhang, et al. 2015. "Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio Parahaemolyticus." *Molecular Biology and Evolution* 32 (6):1396–1410.

Degnan, Patrick H., Howard Ochman, and Nancy A. Moran. 2011. “Sequence Conservation and Functional Constraint on Intergenic Spacers in Reduced Genomes of the Obligate Symbiont Buchnera.” *PLoS Genetics* 7 (9).

De Hoors, Michael J. L., Yuko Makita, Kersta Nakai, and Satora Msyasio. 2005. “Prediction of Transcriptional Terminators in Bacillus Subtiiis and Species.” *PLoS Computational Biology* 1 (3):0212–21.

Desjardins, Christopher A., Keira A. Cohen, Vanisha Munsamy, Thomas Abeel, Kashmeel Maharaj, Bruce J. Walker, Terrance P. Shea, et al. 2016. “Genomic and Functional Analyses of Mycobacterium Tuberculosis Strains Implicate Ald in D-Cycloserine Resistance.” *Nature Genetics* 48 (5):544–51.

Didelot, Xavier, Rory Bowden, Teresa Street, Tanya Golubchik, Chris Spencer, Gil McVean, Vartul Sangal, et al. 2011. “Recombination and Population Structure in Salmonella Enterica.” *PLoS Genetics* 7 (7):e1002191.

Drake, Jared A., Christine Bird, James Nemesh, Daryl J. Thomas, Christopher Newton-Cheh, Alexandre Reymond, Laurent Excoffier, et al. 2006. “Conserved Noncoding Sequences Are Selectively Constrained and Not Mutation Cold Spots.” *Nature Genetics* 38 (2):223–27.

Earle, Sarah G., Chieh-Hsi Wu, Jane Charlesworth, Nicole Stoesser, N. Claire Gordon, Timothy M. Walker, Chris C. A. Spencer, et al. 2016. “Identifying Lineage Effects When Controlling for Population Structure Improves Power in Bacterial Association Studies.” *Nature Microbiology* 1 (April):16041.

Egan, Elizabeth S., Michael A. Fogel, and Matthew K. Waldor. 2005. “Divided Genomes: Negotiating the Cell Cycle in Prokaryotes with Multiple Chromosomes.” *Molecular Microbiology* 56 (5):1129–38.

Enright, M. C., and B. G. Spratt. 1999. “Multilocus Sequence Typing.” *Trends in Microbiology* 7 (12):482–87.

Falush, Daniel, Thierry Wirth, Bodo Linz, Jonathan K. Pritchard, Matthew Stephens, Mark Kidd, Martin J. Blaser, et al. 2003. “Traces of Human Migrations in Helicobacter Pylori Populations.” *Science* 299 (5612):1582–85.

Farhat, Maha R., B. Jesse Shapiro, Karen J. Kieser, Razvan Sultana, Karen R. Jacobson, Thomas C. Victor, Robin M. Warren, et al. 2013. “Genomic Analysis Identifies Targets of Convergent Positive Selection in Drug-Resistant Mycobacterium Tuberculosis.” *Nature Genetics* 45 (10):1183–89.

Feil, Edward J. 2015. “Toward a Synthesis of Genotypic Typing and Phenotypic Inference in the

Genomics Era." *Future Microbiology* 10 (December):1897–99.

Feil, Edward J., Jessica E. Cooper, Hajo Grundmann, D. Ashley Robinson, Mark C. Enright, Tony Berendt, Sharon J. Peacock, et al. 2003. "How Clonal Is Staphylococcus Aureus?" *Journal of Bacteriology* 185 (11):3307–16.

Feil, Edward J., Bao C. Li, David M. Aanensen, William P. Hanage, and Brian G. Spratt. 2004. "eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data." *Journal of Bacteriology* 186 (5):1518–30.

Feiner, Ron, Tal Argov, Lev Rabinovich, Nadejda Sigal, Ilya Borovok, and Anat A. Herskovits. 2015. "A New Perspective on Lysogeny: Prophages as Active Regulatory Switches of Bacteria." *Nature Reviews. Microbiology* 13 (10):641–50.

Feng, Ye, and Cheng-Hsun Chiu. 2014. "Predicting Genetic and Ecological Characteristics of Bacterial Species by Comparing the Trajectories of dN/dS and dI/dS in Bacterial Genomes." *Molecular bioSystems* 10 (2):266–72.

Fishbein, S., N. van Wyk, R. M. Warren, and S. L. Sampson. 2015. "Phylogeny to Function: PE/PPE Protein Evolution and Impact on Mycobacterium Tuberculosis Pathogenicity." *Molecular Microbiology* 96 (5):901–16.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. 1995. "Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd." *Science* 269 (5223):496–512.

Fouts, Derrick E., Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton. 2012. "PanOCT: Automated Clustering of Orthologs Using Conserved Gene Neighborhood for Pan-Genomic Analysis of Bacterial Strains and Closely Related Species." *Nucleic Acids Research* 40 (22):e172.

Frampton, Matthew, and Richard Houlston. 2012. "Generation of Artificial FASTQ Files to Evaluate the Performance of next-Generation Sequencing Pipelines." *PloS One* 7 (11):e49110.

Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, et al. 1997. "Genomic Sequence of a Lyme Disease Spirochaete, Borrelia Burgdorferi." *Nature* 390 (6660):580–86.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data." *Bioinformatics* 28 (23):3150–52.

Fu, Songzhe, Sophie Octavia, Mark M. Tanaka, Vitali Sintchenko, and Ruiting Lan. 2015.

"Defining the Core Genome of Salmonella Enterica Serovar Typhimurium for Genomic Surveillance and Epidemiological Typing." *Journal of Clinical Microbiology* 53 (8):2530–38.

Galardini, Marco, Alexandra Koumoutsi, Lucia Herrera-Dominguez, Juan Antonio Cordero Varela, Anja Telzerow, Omar Wagih, Morgane Wartel, et al. 2017. "Phenotype Prediction in an Escherichia Coli Strain Panel." *bioRxiv*. https://doi.org/10.1101/141879.

Gong, Hao, Gia Phong Vu, Yong Bai, Elton Chan, Ruobin Wu, Edward Yang, Fenyong Liu, and Sangwei Lu. 2011. "A Salmonella Small Non-Coding Rna Facilitates Bacterial Invasion and Intracellular Replication by Modulating the Expression of Virulence Factors." *PLoS Pathogens* 7 (9). http://www.ncbi.nlm.nih.gov/pubmed/21949647.

Gragg, Hana, Brian D. Harfe, and Sue Jinks-Robertson. 2002. "Base Composition of Mononucleotide Runs Affects DNA Polymerase Slippage and Removal of Frameshift Intermediates by Mismatch Repair in Saccharomyces Cerevisiae." *Molecular and Cellular Biology* 22 (24):8756–62.

Guglielmini, Julien, Leonor Quintais, Maria Pilar Garcillán-Barcia, Fernando de la Cruz, and Eduardo P. C. Rocha. 2011. "The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation." *PLoS Genetics* 7 (8):e1002222.

Gu, Tingting, Shengjun Tan, Xiaoxi Gou, Hitoshi Araki, and Dacheng Tian. 2010. "Avoidance of Long Mononucleotide Repeats in Codon Pair Usage." *Genetics* 186 (3):1077–84.

Gweon, Hyun S., Mark J. Bailey, and Daniel S. Read. 2017. "Assessment of the Bimodality in the Distribution of Bacterial Genome Sizes." *The ISME Journal* 11 (3):821–24.

Hadfield, James, Nicholas J. Croucher, Richard J. Goater, Khalil Abudahab, David M. Aanensen, and Simon R. Harris. 2017. "Phandango: An Interactive Viewer for Bacterial Population Genomics." *bioRxiv*. https://doi.org/10.1101/119545.

Hammarlöf, Disa L., Carsten Kröger, Siân V. Owen, Rocío Canals, Lizeth Lacharme-Lora, Nicolas Wenner, Timothy J. Wells, et al. 2017. "The Role of a Single Non-Coding Nucleotide in the Evolution of an Epidemic African Clade of Salmonella." *bioRxiv*. https://doi.org/10.1101/175265.

Han, Kui, Zhi-Feng Li, Ran Peng, Li-Ping Zhu, Tao Zhou, Lu-Guang Wang, Shu-Guang Li, et al. 2013. "Extraordinary Expansion of a Sorangium Cellulosum Genome from an Alkaline Milieu." *Scientific Reports* 3:2101.

Harrison, Ellie, and Michael A. Brockhurst. 2012. "Plasmid-Mediated Horizontal Gene Transfer Is a Coevolutionary Process." *Trends in Microbiology* 20 (6):262–67.

Harris, Simon R., Edward J. Feil, Matthew T. G. Holden, Michael A. Quail, Emma K. Nickerson,

Narisara Chantratita, Susana Gardete, et al. 2010. “Evolution of MRSA during Hospital Transmission and Intercontinental Spread.” *Science* 327 (5964):469–74.

Hartl, D. L., E. N. Moriyama, and S. A. Sawyer. 1994. “Selection Intensity for Codon Bias.” *Genetics* 138 (1):227–34.

Hawley, D. K., and W. R. McClure. 1983. “Compilation and Analysis of Escherichia Coli Promoter DNA Sequences.” *Nucleic Acids Research* 11 (8):2237–55.

Hershberg, Ruth, Mikhail Lipatov, Peter M. Small, Hadar Sheffer, Stefan Niemann, Susanne Homolka, Jared C. Roach, et al. 2008. “High Functional Diversity in Mycobacterium Tuberculosis Driven by Genetic Drift and Human Demography.” *PLoS Biology* 6 (12):e311.

Hershberg, Ruth, and Dmitri A. Petrov. 2010. “Evidence That Mutation Is Universally Biased towards AT in Bacteria.” *PLoS Genetics* 6 (9):e1001115.

Hildebrand, Falk, Axel Meyer, and Adam Eyre-Walker. 2010. “Evidence of Selection upon Genomic GC-Content in Bacteria.” *PLoS Genetics* 6 (9).

Holden, Matthew T. G., Li-Yang Hsu, Kevin Kurt, Lucy A. Weinert, Alison E. Mather, Simon R. Harris, Birgit Strommenger, et al. 2013. “A Genomic Portrait of the Emergence, Evolution, and Global Spread of a Methicillin-Resistant Staphylococcus Aureus Pandemic.” *Genome Research* 23 (4):653–64.

Holt, Kathryn E., Heiman Wertheim, Ruth N. Zadoks, Stephen Baker, Chris A. Whitehouse, David Dance, Adam Jenney, et al. 2015. “Genomic Analysis of Diversity, Population Structure, Virulence, and Antimicrobial Resistance in Klebsiella Pneumoniae, an Urgent Threat to Public Health.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (27):E3574–81.

Hook-Barnard, India G., and Deborah M. Hinton. 2007. “Transcription Initiation by Mix and Match Elements: Flexibility for Polymerase Binding to Bacterial Promoters.” *Gene Regulation and Systems Biology* 1:275–93.

Hu, Honghua, Ruiting Lan, and Peter R. Reeves. 2006. “Adaptation of Multilocus Sequencing for Studying Variation within a Major Clone: Evolutionary Relationships of Salmonella Enterica Serovar Typhimurium.” *Genetics* 172 (2):743–50.

Hurst, Laurence D. 2002. “The Ka/Ks Ratio: Diagnosing the Form of Sequence Evolution.” *Trends in Genetics: TIG* 18 (9):486.

Jacob, F., and J. Monod. 1961. “Genetic Regulatory Mechanisms in the Synthesis of Proteins.” *Journal of Molecular Biology* 3 (June):318–56.

Johnston, Calum, Bernard Martin, Gwennaele Fichant, Patrice Polard, and Jean-Pierre

Claverys. 2014. “Bacterial Transformation: Distribution, Shared Mechanisms and Divergent Control.” *Nature Reviews. Microbiology* 12 (3):181–96.

Jolley, Keith A., Carly M. Bliss, Julia S. Bennett, Holly B. Bratcher, Carina Brehony, Frances M. Colles, Helen Wimalarathna, et al. 2012. “Ribosomal Multilocus Sequence Typing: Universal Characterization of Bacteria from Domain to Strain.” *Microbiology* 158 (Pt 4):1005–15.

Jolley, Keith A., and Martin C. J. Maiden. 2010. “BIGSdb: Scalable Analysis of Bacterial Genome Variation at the Population Level.” *BMC Bioinformatics* 11 (1):595.

Jong, Anne de, Hilco Pietersma, Martijn Cordes, Oscar P. Kuipers, and Jan Kok. 2012. “PePPER: A Webserver for Prediction of Prokaryote Promoter Elements and Regulons.” *BMC Genomics* 13 (July):299.

Juhas, Mario, Derrick W. Crook, and Derek W. Hood. 2008. “Type IV Secretion Systems: Tools of Bacterial Horizontal Gene Transfer and Virulence.” *Cellular Microbiology* 10 (12):2377–86.

Jukes, Thomas H., and Charles R. Cantor. 1969. “CHAPTER 24 - Evolution of Protein Molecules A2 - MUNRO, H.N.” In *Mammalian Protein Metabolism*, 21–132. Academic Press.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4):772–80.

Kimura, M. 1991. “The Neutral Theory of Molecular Evolution: A Review of Recent Evidence.” *Idengaku Zasshi* 66 (4):367–86.

Kimura, M., and T. Ohta. 1971. “Protein Polymorphism as a Phase of Molecular Evolution.” *Nature* 229 (5285):467–69.

Kingsford, Carleton L., Kunmi Ayanbule, and Steven L. Salzberg. 2007. “Rapid, Accurate, Computational Discovery of Rho-Independent Transcription Terminators Illuminates Their Relationship to DNA Uptake.” *Genome Biology* 8 (2):R22.

Kodaman, Nuri, Alvaro Pazos, Barbara G. Schneider, M. Blanca Piazuelo, Robertino Mera, Rafal S. Sobota, Liviu A. Sicinschi, et al. 2014. “Human and Helicobacter Pylori Coevolution Shapes the Risk of Gastric Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (4):1455–60.

Koonin, Eugene V., and Yuri I. Wolf. 2008. “Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World.” *Nucleic Acids Research* 36 (21):6688–6719.

Laabei, Maisem, Mario Recker, Justine K. Rudkin, Mona Aldeljawi, Zeynep Gulay, Tim J. Sloan, Paul Williams, et al. 2014. “Predicting the Virulence of MRSA from Its Genome Sequence.” *Genome Research* 24 (5):839–49.

Larsson, Christer, Brian Luna, Nicole C. Ammerman, Mamoudou Maiga, Nisheeth Agarwal, and William R. Bishai. 2012. “Gene Expression of Mycobacterium Tuberculosis Putative Transcription Factors whiB1-7 in Redox Environments.” *PloS One* 7 (7):e37516.

Lawson, Daniel John, Garrett Hellenthal, Simon Myers, and Daniel Falush. 2012. “Inference of Population Structure Using Dense Haplotype Data.” *PLoS Genetics* 8 (1):e1002453.

Lees, John A., Minna Vehkala, Niko Välimäki, Simon R. Harris, Claire Chewapreecha, Nicholas J. Croucher, Pekka Marttinen, et al. 2016. “Sequence Element Enrichment Analysis to Determine the Genetic Basis of Bacterial Phenotypes.” *Nature Communications* 7 (September):12797.

Lewis, M., G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu. 1996. “Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer.” *Science* 271 (5253):1247–54.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16):2078–79.

Lin, Wei Hsiang, and Edo Kussell. 2012. “Evolutionary Pressures on Simple Sequence Repeats in Prokaryotic Coding Regions.” *Nucleic Acids Research* 40 (6):2399–2413.

Linz, Bodo, François Balloux, Yoshan Moodley, Andrea Manica, Hua Liu, Philippe Roumagnac, Daniel Falush, et al. 2007. “An African Origin for the Intimate Association between Humans and Helicobacter Pylori.” *Nature* 445 (7130):915–18.

Luo, Haiwei, Jijun Tang, Robert Friedman, and Austin L. Hughes. 2011. “Ongoing Purifying Selection on Intergenic Spacers in Group A Streptococcus.” *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 11 (2):343–48.

Macfarlane, L., J. Walker, R. Borrow, B. A. Oppenheim, and A. J. Fox. 1999. “Improved Recognition of MRSA Case Clusters by the Application of Molecular Subtyping Using Pulsed-Field Gel Electrophoresis.” *The Journal of Hospital Infection* 41 (1):29–37.

Maiden, Martin C. J., and Odile B. Harrison. 2016. “The Population and Functional Genomics of the Neisseria Revealed with Gene-by-Gene Approaches.” *Journal of Clinical Microbiology*,

April. https://doi.org/10.1128/JCM.00301-16.

Maiden, Martin C. J., Melissa J. Jansen van Rensburg, James E. Bray, Sarah G. Earle, Suzanne A. Ford, Keith A. Jolley, and Noel D. McCarthy. 2013. “MLST Revisited: The Gene-by-Gene Approach to Bacterial Genomics.” *Nature Reviews. Microbiology* 11 (10):728–36.

Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, et al. 1998. “Multilocus Sequence Typing: A Portable Approach to the Identification of Clones within Populations of Pathogenic Microorganisms.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (6):3140–45.

Maixner, Frank, Ben Krause-Kyora, Dmitrij Turaev, Alexander Herbig, Michael R. Hoopmann, Janice L. Hallows, Ulrike Kusebauch, et al. 2016. “The 5300-Year-Old Helicobacter Pylori Genome of the Iceman.” *Science* 351 (6269):162–65.

Ma, Shuyi, Kyle J. Minch, Tige R. Rustad, Samuel Hobbs, Suk-Lin Zhou, David R. Sherman, and Nathan D. Price. 2015. “Integrated Modeling of Gene Regulatory and Metabolic Networks in Mycobacterium Tuberculosis.” *PLoS Computational Biology* 11 (11):e1004543.

McCutcheon, John P., and Nancy A. Moran. 2011. “Extreme Genome Reduction in Symbiotic Bacteria.” *Nature Reviews. Microbiology* 10 (1):13–26.

McInerney, James O., Alan McNally, and Mary J. O’Connell. 2017. “Why Prokaryotes Have Pangenomes.” *Nature Microbiology* 2 (March):17040.

McNally, Alan, Yaara Oren, Darren Kelly, Ben Pascoe, Steven Dunn, Tristan Sreecharan, Minna Vehkala, et al. 2016. “Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations.” *PLoS Genetics* 12 (9):e1006280.

Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. 2005. “The Microbial Pan-Genome.” *Current Opinion in Genetics & Development* 15 (6):589–94.

Milkman, R. 1973. “Electrophoretic Variation in Escherichia Coli from Natural Sources.” *Science* 182 (4116):1024–26.

Molina, Nacho, and Erik Van Nimwegen. 2008. “Universal Patterns of Purifying Selection at Noncoding Positions in Bacteria.” *Genome Research* 18 (1):148–60.

Moodley, Yoshan, Bodo Linz, Robert P. Bond, Martin Nieuwoudt, Himla Soodyall, Carina M. Schlebusch, Steffi Bernhöft, et al. 2012. “Age of the Association between Helicobacter Pylori and Man.” *PLoS Pathogens* 8 (5):e1002693.

Moran, N. A., and A. Mira. 2001. “The Process of Genome Shrinkage in the Obligate Symbiont

Buchnera Aphidicola." *Genome Biology* 2 (12):RESEARCH0054.

Muto, A., and S. Osawa. 1987. "The Guanine and Cytosine Content of Genomic DNA and Bacterial Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 84 (1):166–69.

Nakabachi, Atsushi, Atsushi Yamashita, Hidehiro Toh, Hajime Ishikawa, Helen E. Dunbar, Nancy A. Moran, and Masahira Hattori. 2006. "The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella." *Science* 314 (5797):267.

Namouchi, Amine, Xavier Didelot, Ulrike Schöck, Brigitte Gicquel, and Eduardo P. C. Rocha. 2012. "After the Bottleneck: Genome-Wide Diversification of the Mycobacterium Tuberculosis Complex by Mutation, Recombination, and Natural Selection." *Genome Research* 22 (4):721–34.

Nei, M., and T. Gojobori. 1986. "Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions." *Molecular Biology and Evolution* 3 (5):418–26.

Niehus, Rene, Sara Mitri, Alexander G. Fletcher, and Kevin R. Foster. 2015. "Migration and Horizontal Gene Transfer Divide Microbial Genomes into Multiple Niches." *Nature Communications* 6 (November):8924.

Ohta, T. 1973. "Slightly Deleterious Mutant Substitutions in Evolution." *Nature* 246 (5428):96–98.

Oren, Yaara, Mark B. Smith, Nathan I. Johns, Millie Kaplan Zeevi, Dvora Biran, Eliora Z. Ron, Jukka Corander, Harris H. Wang, Eric J. Alm, and Tal Pupko. 2014. "Transfer of Noncoding DNA Drives Regulatory Rewiring in Bacteria." *Proceedings of the National Academy of Sciences of the United States of America* 111 (45):16112–17.

Osório, Nuno S., Fernando Rodrigues, Sebastien Gagneux, Jorge Pedrosa, Marta Pinto-Carbó, António G. Castro, Douglas Young, Iñaki Comas, and Margarida Saraiva. 2013. "Evidence for Diversifying Selection in a Set of Mycobacterium Tuberculosis Genes in Response to Antibiotic- and Nonantibiotic-Related Pressure." *Molecular Biology and Evolution* 30 (6):1326–36.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. "Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis." *Bioinformatics* 31 (22):3691–93.

Penadés, José R., John Chen, Nuria Quiles-Puchalt, Nuria Carpena, and Richard P. Novick.

2015. “Bacteriophage-Mediated Spread of Bacterial Virulence Genes.” *Current Opinion in Microbiology* 23 (February):171–78.

Pepperell, Caitlin S., Amanda M. Casto, Andrew Kitchen, Julie M. Granka, Omar E. Cornejo, Edward C. Holmes, Eddie C. Holmes, Bruce Birren, James Galagan, and Marcus W. Feldman. 2013. “The Role of Selection in Shaping Diversity of Natural M. Tuberculosis Populations.” *PLoS Pathogens* 9 (8):e1003543.

Peters, Jason M., Abbey D. Vangeloff, and Robert Landick. 2011. “Bacterial Transcription Terminators: The RNA 3’-End Chronicles.” *Journal of Molecular Biology* 412 (5):793–813.

Pimentel, Harold, Nicolas L. Bray, Suzette Puente, Páll Melsted, and Lior Pachter. 2017. “Differential Analysis of RNA-Seq Incorporating Quantification Uncertainty.” *Nature Methods* 14 (7):687–90.

Price, Morgan N., and Adam P. Arkin. 2015. “Weakly Deleterious Mutations and Low Rates of Recombination Limit the Impact of Natural Selection on Bacterial Genomes.” *mBio* 6 (6). https://doi.org/10.1128/mBio.01302-15.

Raghavan, Rahul, Yogeshwar D. Kelkar, and Howard Ochman. 2012. “A Selective Force Favoring Increased G+C Content in Bacterial Genes.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (36):14504–7.

Raimunda, Daniel, Jarukit E. Long, Teresita Padilla-Benavides, Christopher M. Sassetti, and José M. Argüello. 2014. “Differential Roles for the Co(2+) /Ni(2+) Transporting ATPases, CtpD and CtpJ, in Mycobacterium Tuberculosis Virulence.” *Molecular Microbiology* 91 (1):185–97.

Reuter, Sandra, Estee M. Török, Matthew T. G. Holden, Rosy Reynolds, Kathy E. Raven, Beth Blane, Tjibbe Donker, et al. 2015. “Building a Genomic Framework for Prospective MRSA Surveillance in the United Kingdom and the Republic of Ireland.” *Genome Research*, December. https://doi.org/10.1101/gr.196709.115.

Roberts, R. B., A. de Lencastre, W. Eisner, E. P. Severina, B. Shopsin, B. N. Kreiswirth, and A. Tomasz. 1998. “Molecular Epidemiology of Methicillin-Resistant Staphylococcus Aureus in 12 New York Hospitals. MRSA Collaborative Study Group.” *The Journal of Infectious Diseases* 178 (1):164–71.

Rocha, Eduardo P. C., and Edward J. Feil. 2010. “Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria?” *PLoS Genetics* 6 (9).

Rocha, Eduardo P. C., John Maynard Smith, Laurence D. Hurst, Matthew T. G. Holden, Jessica

E. Cooper, Noel H. Smith, and Edward J. Feil. 2006. “Comparisons of dN/dS Are Time Dependent for Closely Related Bacterial Genomes.” *Journal of Theoretical Biology* 239 (2):226–35.

Rohwer, Forest, and Rob Edwards. 2002. “The Phage Proteomic Tree: A Genome-Based Taxonomy for Phage.” *Journal of Bacteriology* 184 (16):4529–35.

Romilly, Cédric, Claire Lays, Arnaud Tomasini, Isabelle Caldelari, Yvonne Benito, Philippe Hammann, Thomas Geissmann, Sandrine Boisset, Pascale Romby, and François Vandenesch. 2014. “A Non-Coding RNA Promotes Bacterial Persistence and Decreases Virulence by Regulating a Regulator in Staphylococcus Aureus.” *PLoS Pathogens* 10 (3). Public Library of Science. http://www.ncbi.nlm.nih.gov/pubmed/24651379.

Safina, Ksenia R., Andrey A. Mironov, and Georgii A. Bazykin. 2017. “Compensatory Evolution of Intrinsic Transcription Terminators in Bacillus Cereus.” *Genome Biology and Evolution*, February. https://doi.org/10.1093/gbe/evw295.

Sahl, Jason W., J. Gregory Caporaso, David A. Rasko, and Paul Keim. 2014. “The Large-Scale Blast Score Ratio (LS-BSR) Pipeline: A Method to Rapidly Compare Genetic Content between Bacterial Genomes.” *PeerJ* 2 (April):e332.

Santangelo, Thomas J., and Irina Artsimovitch. 2011. “Termination and Antitermination: RNA Polymerase Runs a Stop Sign.” *Nature Reviews. Microbiology* 9 (5):319–29.

Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14):2068–69.

Selander, R. K., D. A. Caugant, H. Ochman, J. M. Musser, M. N. Gilmour, and T. S. Whittam. 1986. “Methods of Multilocus Enzyme Electrophoresis for Bacterial Population Genetics and Systematics.” *Applied and Environmental Microbiology* 51 (5):873–84.

Sharp, Paul M., Elizabeth Bailes, Russell J. Grocock, John F. Peden, and R. Elizabeth Sockett. 2005. “Variation in the Strength of Selected Codon Usage Bias among Bacteria.” *Nucleic Acids Research* 33 (4):1141–53.

Sheppard, Samuel K., Xavier Didelot, Guillaume Meric, Alicia Torralbo, Keith A. Jolley, David J. Kelly, Stephen D. Bentley, Martin C. J. Maiden, Julian Parkhill, and Daniel Falush. 2013. “Genome-Wide Association Study Identifies Vitamin B5 Biosynthesis as a Host Specificity Factor in Campylobacter.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (29):11923–27.

Sheppard, Samuel K., Keith A. Jolley, and Martin C. J. Maiden. 2012. “A Gene-by-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter.”

Genes 3 (2):261–77.

Shimada, Tomohiro, Yukiko Yamazaki, Kan Tanaka, and Akira Ishihama. 2014. “The Whole Set of Constitutive Promoters Recognized by RNA Polymerase RpoD Holoenzyme of Escherichia Coli.” *PloS One* 9 (3):e90447.

Sirakova, T. D., A. K. Thirumala, V. S. Dubey, H. Sprecher, and P. E. Kolattukudy. 2001. “The Mycobacterium Tuberculosis pks2 Gene Encodes the Synthase for the Hepta- and Octamethyl-Branched Fatty Acids Required for Sulfolipid Synthesis.” *The Journal of Biological Chemistry* 276 (20):16833–39.

Skwark, Marcin J., Nicholas J. Croucher, Santeri Puranen, Claire Chewapreecha, Maiju Pesonen, Ying Ying Xu, Paul Turner, et al. 2017. “Interacting Networks of Resistance, Virulence and Core Machinery Genes Identified by Genome-Wide Epistasis Analysis.” *PLoS Genetics* 13 (2):e1006508.

Smith, J. M., N. H. Smith, M. O’Rourke, and B. G. Spratt. 1993. “How Clonal Are Bacteria?” *Proceedings of the National Academy of Sciences of the United States of America* 90 (10):4384–88.

Smith, Laura J., Melanie R. Stapleton, Roger S. Buxton, and Jeffrey Green. 2012. “Structure-Function Relationships of the Mycobacterium Tuberculosis Transcription Factor WhiB1.” *PloS One* 7 (7):e40407.

Stranger, Barbara E., Eli A. Stahl, and Towfique Raj. 2011. “Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics.” *Genetics* 187 (2):367–83.

Suwanto, A., and S. Kaplan. 1989. “Physical and Genetic Mapping of the Rhodobacter Sphaeroides 2.4.1 Genome: Presence of Two Unique Circular Chromosomes.” *Journal of Bacteriology* 171 (11):5850–59.

Suzek, B. E., M. D. Ermolaeva, M. Schreiber, and S. L. Salzberg. 2001. “A Probabilistic Method for Identifying Start Codons in Bacterial Genomes.” *Bioinformatics* 17 (12):1123–30.

Thorell, Kaisa, Koji Yahara, Elvire Berthenet, Daniel J. Lawson, Ikuko Kato, Alfonso Tenorio Mendez, Federico Canzian, et al. 2016. “Rapid Evolution of Distinct Helicobacter Pylori Subpopulations in the Americas.” *bioRxiv*. https://doi.org/10.1101/069070.

Thorpe, Harry A., Sion C. Bayliss, Laurence D. Hurst, and Edward J. Feil. 2017. “Comparative Analyses of Selection Operating on Non-Translated Intergenic Regions of Diverse Bacterial Species.” *Genetics*, March. https://doi.org/10.1534/genetics.116.195784.

Tjaden, Brian. 2015. “De Novo Assembly of Bacterial Transcriptomes from RNA-Seq Data.”

Genome Biology 16 (January):1.

Vale, Pedro F., and Tom J. Little. 2010. “CRISPR-Mediated Phage Resistance and the Ghost of Coevolution Past.” *Proceedings. Biological Sciences / The Royal Society* 277 (1691):2097–2103.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. “The Sequence of the Human Genome.” *Science* 291 (5507):1304–51.

Vos, Michiel, and Xavier Didelot. 2009. “A Comparison of Homologous Recombination Rates in Bacteria and Archaea.” *The ISME Journal* 3 (2):199–208.

Vos, Michiel, Matthijn C. Hesselman, Tim A. te Beek, Mark W. J. van Passel, and Adam Eyre-Walker. 2015. “Rates of Lateral Gene Transfer in Prokaryotes: High but Why?” *Trends in Microbiology* 23 (10):598–605.

Wang, Tai-Chun, and Feng-Chi Chen. 2013. “The Evolutionary Landscape of the Mycobacterium Tuberculosis Genome.” *Gene* 518 (1):187–93.

Warne, Ben, Catriona P. Harkins, Simon R. Harris, Alexandra Vatsiou, Nicola Stanley-Wall, Julian Parkhill, Sharon J. Peacock, Tracy Palmer, and Matthew T. G. Holden. 2016. “The Ess/Type VII Secretion System of Staphylococcus Aureus Shows Unexpected Genetic Diversity.” *BMC Genomics* 17 (March):222.

Washburn, R. S., A. Marra, A. P. Bryant, M. Rosenberg, and D. R. Gentry. 2001. “Rho Is Not Essential for Viability or Virulence in Staphylococcus Aureus.” *Antimicrobial Agents and Chemotherapy* 45 (4):1099–1103.

Washio, T., J. Sasayama, and M. Tomita. 1998. “Analysis of Complete Genomes Suggests That Many Prokaryotes Do Not Rely on Hairpin Formation in Transcription Termination.” *Nucleic Acids Research* 26 (23):5456–63.

Waters, Lauren S., and Gisela Storz. 2009. “Regulatory RNAs in Bacteria.” *Cell* 136 (4):615–28.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer Publishing Company, Incorporated.

Wigneshweraraj, Sivaramesh, Daniel Bose, Patricia C. Burrows, Nicolas Joly, Jörg Schumacher, Mathieu Rappas, Tillmann Pape, et al. 2008. “Modus Operandi of the Bacterial RNA Polymerase Containing the sigma54 Promoter-Specificity Factor.” *Molecular Microbiology* 68 (3):538–46.

Yahara, Koji, Yoshikazu Furuta, Kenshiro Oshima, Masaru Yoshida, Takeshi Azuma, Masahira Hattori, Ikuo Uchiyama, and Ichizo Kobayashi. 2013. “Chromosome Painting in Silico in a Bacterial Species Reveals Fine Population Structure.” *Molecular Biology and Evolution* 30

(6):1454–64.

Yang, Ziheng. 2007. “PAML 4: Phylogenetic Analysis by Maximum Likelihood.” *Molecular Biology and Evolution* 24 (8):1586–91.

Yang, Z., and R. Nielsen. 2000. “Estimating Synonymous and Nonsynonymous Substitution Rates under Realistic Evolutionary Models.” *Molecular Biology and Evolution* 17 (1):32–43.

Zhao, Yongbing, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. 2012. “PGAP: Pan-Genomes Analysis Pipeline.” *Bioinformatics* 28 (3):416–18.

Zhou, Zhemin, Angela McCann, François-Xavier Weill, Camille Blin, Satheesh Nair, John Wain, Gordon Dougan, and Mark Achtman. 2014. “Transient Darwinian Selection in Salmonella Enterica Serovar Paratyphi A during 450 Years of Global Spread of Enteric Fever.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (33):12199–204.